DOCUMENT RESUME

ED 251 425 SP 025 433

AUTHOR Rogosa, David; And Others
TITLE Assessing the Stability of Teacher Behavior. Research Series No. 141.
INSTITUTION Michigan State Univ., East Lansing. Inst. for Research on Teaching.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Apr 84
CONTRACT 400-81-0014
NOTE 78p.
AVAILABLE FROM Institute for Research on Teaching, College of Education, Michigan State University, 252 Erickson Hall, E. Lansing, MI 48824 ($8.00).
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Conceptual Tempo; Elementary Secondary Education; *Individual Differences; Personality Measures; Personality Traits; *Psychological Characteristics; *Reliability; *Research Methodology; *Teacher Behavior; *Teacher Characteristics

ABSTRACT

A study of the stability of teacher behavior over time was formulated through two major questions: (1) Is the behavior of an individual teacher consistent over time? and (2) Are individual differences among teachers consistent over time? Regrettably, the first question has rarely been considered in previous investigations of the stability of teacher behavior, and empirical research on the second question has been marked by considerable confusion. In this paper, statistical procedures have been developed for answering each of these questions. Approaches and methods of previous studies of temporal stability are re-evaluated. In addition, methods for assessing the stability of teacher behavior across contexts are described. Observational data on classroom teachers are used throughout to illustrate new approaches and methods for the study of stability of behavior. (Author/JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

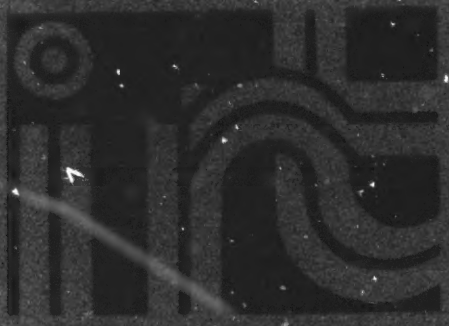
Robert E. Floden

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Research Series No. 141

ASSESSING THE STABILITY OF TEACHER BEHAVIOR

David Rogosa, Robert E. Floden,
and John B. Willett



Institute
for
Research on Teaching

College of Education Michigan State University

Research Series No. 141

ASSESSING THE STABILITY OF TEACHER BEHAVIOR

David Rogosa, Robert E. Floden,
and John B. Willett

Published By

The Institute for Research on Teaching
252 Erickson Hall
Michigan State University
East Lansing, Michigan 48824

April 1984

Publication of this work is sponsored by the Institute for Research on Teaching, College of Education, Michigan State University. The Institute for Research on Teaching is funded primarily by the Program for Teaching and Instruction of the National Institute of Education, United States Department of Education. The opinions expressed in this publication do not necessarily reflect the position, policy, or endorsement of the National Institute of Education. (Contract No. 400-81-0014)

Institute for Research on Teaching

The **Institute for Research on Teaching** was founded at Michigan State University in 1976 by the National Institute of Education. Following a nationwide competition in 1981, the NIE awarded a second contract to the IRT, extending work through 1984. Funding is also received from other agencies and foundations for individual research projects.

The IRT conducts major research projects aimed at improving classroom teaching, including studies of classroom management strategies, student socialization, the diagnosis and remediation of reading difficulties, and teacher education. IRT researchers are also examining the teaching of specific school subjects such as reading, writing, general mathematics, and science, and are seeking to understand how factors outside the classroom affect teacher decision making.

Researchers from such diverse disciplines as educational psychology, anthropology, sociology, and philosophy cooperate in conducting IRT research. They join forces with public school teachers, who work at the IRT as half-time collaborators in research, helping to design and plan studies, collect data, analyze and interpret results, and disseminate findings.

The IRT publishes research reports, occasional papers, conference proceedings, a newsletter for practitioners, and lists and catalogs of IRT publications. For more information, to receive a list or catalog, and/or to be placed on the IRT mailing list to receive the newsletter, please write to the IRT Editor, Institute for Research on Teaching, 252 Erickson Hall, Michigan State University, East Lansing, Michigan 48824-1034.

Co-Directors: Jere E. Brophy and Andrew C. Porter

Associate Directors: Judith E. Lanier and Richard S. Prawat

Editorial Staff
Editor: Janet Eaton
Assistant Editor: Patricia Nischan

Abstract

This study of the stability of teacher behavior over time is formulated through two major questions: (1) Is the behavior of an individual teacher consistent over time? and (2) Are individual differences among teachers consistent over time? Regrettably, the first question has rarely been considered in previous investigations of the stability of teacher behavior, and empirical research on the second question has been marked by considerable confusion. In this paper we develop statistical procedures for answering each of these questions. Approaches and methods of previous studies of temporal stability are re-evaluated. In addition, methods for assessing the stability of teacher behavior across contexts are described. Observational data on classroom teachers are used throughout to illustrate our new approaches and methods for the study of stability of behavior.

ASSESSING THE STABILITY OF TEACHER BEHAVIOR[1]

David Rogosa, Robert E. Floden, and John B. Willett[2]

Questions About Stability

Answers to the question, "Is teacher behavior stable?" have been sought by researchers for the past half-century. Despite the many extensive studies of teacher behavior, affirmative answers to this question have been rare. For example, Borich (1977) states, "The results of these studies suggest that teacher behavior may be unstable across long periods of time and content" (p. 300)[3]. Moreover, researchers lack confidence in the results of these research efforts; the conclusion of Shavelson and Dempsey-Atwood (1976) that "(a) most

---

[1]This research was supported primarily by a grant from the Program for Teaching and Instruction of the National Institute of Education, United States Department of Education. (Grant #NIE-G-81-0087) In addition, the work of the first and third authors was supported in part by grants from the Spencer Foundation; the work of the second author was supported in part by the Institute for Research on Teaching, College of Education, Michigan State University.

[2]David Rogosa is an Assistant Professor at the Stanford University School of Education. Robert E. Floden is an IRT senior researcher and an MSU associate professor of teacher education and educational psychology. John B. Willett is a graduate student at the Stanford University School of Education.

The authors wish to thank Donald Veldman of the University of Texas at Austin and Thomas Moon of California State College in Pennsylvania for providing them with the classroom observation data used in this paper.

[3]The recent review by Darling-Hammond, Wise, and Pease (1983) summarizes the research on stability of teaching behavior as follows: "The bottom-line question is, Does a given teacher exhibit the same kinds of behavior at different points in time and within different teaching contexts? In general, the answer is 'no,' especially with regard to measures of specific, discrete teaching behaviors" (p. 299).

studies are methodologically inadequate at this point in time to resolve the issue, (b) findings on the . . . stability of measures of teacher behavior are equivocal with only a few exceptions" (p. 609) is typical of the reviews of this literature.

A major problem in past investigations of stability is that the question, "Is teacher behavior stable?" is not well formulated. Before satisfactory conclusions about stability of teacher behavior can be drawn, the research question must be refined.

In this paper we formulate and pursue two research questions about stability of teacher behavior over time:

Question A. Is the behavior of an individual teacher consistent over time?

Question B. Are individual differences among teachers consistent over time?

Although there have been a few studies of the "variety" and "flexibility" of teacher behavior (see the studies reviewed in Rosenshine, 1971, Chap. 4), *Question A* has received almost no attention in research on teaching. This unfortunate omission reflects the neglect in research on teaching of the detailed description of individual teachers. Virtually every empirical study of stability of teacher behavior has been directed toward answering *Question B*. These studies, however, have not met with great success in demonstrating consistency of individual differences among teachers.

Purposes for Studying Stability

Before turning to the statistical models and methods that we develop for addressing *Questions A* and *B*, it is useful to consider how these questions about stability fit in with current and prior research on teaching. Thus we examine

the purposes for assessing stability of teacher behavior expressed in the research-on-teaching literature. Furthermore, we advocate additional uses for assessments of stability of teacher behavior.

Process-Product Research

Just as one question (*Question B*) has received virtually all the attention in the research literature, one concern has dominated the study of stability of teacher behavior--low stability as a potential obstacle to the establishment of process-product relationships. To demonstrate such relationships, affirmative answers to both *Question A* and *Question B* are crucial. In the research-on-teaching literature, a lack of stability is frequently cited as an explanation for difficulties in demonstrating strong process-product relations. Doyle (1977) notes the importance of stability in process-product research: "If there is wide variability in either the behavior of teachers or the instruments used to measure that behavior, then estimates of process-product relationships are precarious at best" (p. 169). Furthermore, it has been suggested that "measures of teacher behavior may be too unstable to yield consistent relationships with student outcomes" (Shavelson & Dempsey-Atwood, 1976, p.544).

In detecting process-product relationships, the ability to rank teachers reliably on process is essential. Hence, consistent individual differences in teacher behavior facilitate detection of a link between individual differences in teacher behavior and individual differences in teacher effectiveness. Consequently, the effect of negative answers to *Question B* on the magnitudes of potential correlations between process and product measures has been a major concern in research on teacher effectiveness (see Brophy, 1979).

In addition, the interpretability of process-product correlations depends

on an affirmative answer to *Question A*. In process-product correlation strategies, process (teacher behavior) is represented by a single number (often the average over multiple occasions of observation). Medley (1982) explains that the consequence of relying on an average measure of behavior is that "any intentional variations that a teacher introduces to adapt his or her behavior to different purposes are treated as errors of measurement" (p. 1898). Furthermore, an individual teacher's behavior must be consistent over time for a single-number summary of process (i.e., the average over occasions) to be a reasonably complete description of that teacher's activity in the classroom. McGaw, Wardrop, and Bunda (1972) caution that "efforts to develop indices to characterize particular teachers appear to be misplaced unless there is some allowance made for lawful adaptations of behavior to different situations" (p. 16). Other investigators have expressed similar concerns about process-product research: Berliner (1976) requires that "the behavior should be representative of the teacher's *usual* and *customary* way of behaving" (p. 7, emphasis in original) (see also Doyle, 1977, p. 169; Medley & Mitzel, 1963; Medley, 1979, p. 14). Because almost no empirical research on the behavior of individual teachers exists, it is impossible to judge whether the presumption of consistency is reasonable.

We do not mean to imply that considerations of stability should dictate which teacher behaviors are studied. In particular, it is important to note that consistency over time is not necessarily a quality to be prized in teachers. As Berliner (1976) writes:

> Usually people think of good teachers as flexible. Such teachers are expected to change methods, techniques and styles to suit particular students, curriculum areas, time of day or year, etc. That is, the standard of excellence in teaching commonly held implies a teacher whose behavior is inherently unstable. (p. 9)

Flanders (1969, 1970) also prizes teacher flexibility and uses measures of teacher flexibility as a process variable.

Moreover, in understanding effective teaching, variables that are stable are not necessarily important: "There is little reason to presume on a priori grounds that behaviors which are either stable or generalizable across settings are necessarily those that are the most powerful correlates of achievement in a given classroom situation" (Doyle, 1977, p. 169).

Describing Teaching

Interest in stability of teacher behavior as a description of teacher activities has been rare. Yet answers to *Questions A* and *B* are an important part of understanding teaching. As Berliner (1976) asserts, "Until more is known about which teacher behaviors fluctuate, and how and why they fluctuate over time, settings, curricula, and populations, studies relating teacher behavior to student outcomes must remain primitive" (p. 9).

Question A addresses the consistency over time of individual teachers' behavior. The assessment of stability for individual teachers can be used to address many research questions. Characteristics of a group of teachers can be investigated using assessments of the consistency of each teacher's behavior. For a group of teachers, it would be profitable to investigate questions about (a) the group as a whole and (b) individual differences in the consistency over time among group members. Examples of questions about the group as a whole are, "Are most teachers consistent over time in their classroom behavior?" "For which behaviors, if any, are most teachers consistent over time and for which behaviors are most teachers not consistent?" The major question about individual differences in consistency is, "What kinds of teachers are (or are

not) consistent in their behavior?"[4] The detection of systematic individual differences in consistency would be a first step toward understanding how and why teacher behaviors fluctuate over time.

With *Question B*, the focus shifts to descriptions of the consistency of individual differences in teacher behavior over time. (We emphasize that consistency of individual differences is distinct from consistency of an individual teacher and from individual differences in consistency.) Some useful research questions would be "Are individual differences in behavior among teachers consistent over time?" "For what groups of teachers and what behaviors?" "In what situations?" and "Over what period of time?"

Data on a Target Behavior

In this paper we develop methods for assessing stability of behavior using the kinds of data usually available from classroom observation instruments. Hence, an important first step in assessing stability is to understand the structure of the data that are collected. (Lack of attention to the structure of the data collected has weakened the value of many previous studies of stability.) Typically, data used to describe a teacher behavior have been one of three types: (a) behavior-count data, (b) Bernoulli-trial data, and (c) quantitative measures. Here, we develop appropriate statistical models and procedures for each type.

---

[4] In one of the very few empirical studies of consistency of behavior for individual teachers, Flanders (1969) investigated the influence on student achievement of "flexibility" of individual teachers over time and over contexts. Flanders found indications of an association between flexibility and achievement, concluding that "investigations in this area are likely to be rewarding" (p. 109). Thus a possible implication of Flanders' study is that effective teachers may be the ones who are not consistent in their behavior.

Behavior-Count Data

Behavior-count data are a familiar form of classroom observation data. For each occasion of observation, the behavior-counts consist of the number of times a target behavior occurs while observation is in progress. Ordinarily, no information on the timing or the duration of the behaviors is recorded, only the frequency of occurrence of the target behavior.[5] Table 1 shows behavior-count data for two different teachers from the Texas Junior High School Study (Evertson, Anderson, Edgar, Minter, & Brophy, 1977; see Appendix B for a more complete description of these data).

In Table 1 we retain the original teacher identification codes from the Texas Junior High School Study. Shown in the first row for each teacher are counts of behavioral criticism by the junior high school teacher during English instruction. (That is, the target behavior occurred whenever the junior high school English teacher gave a negative evaluation of student behavior, such as, "expressing anger or personal criticism," Evertson & Veldman, 1981, p. 157.) In the second row for each teacher are the number of hours of classroom observation in each month.

To set notation for behavior-count data, let X_i denote the number of recorded occurrences of the target behavior during the observation of an individual teacher on Occasion i. (The subscript i indexes the observation Occasion; i = 1, . . . , T.) Also let b_i indicate the length of the observation period on Occasion i (in Table 1, the number of hours of instruction observed

---

[5]As is discussed in the section on design and in Appendix A, the usual procedure of recording only the frequency of occurrence of the target behavior discards valuable information.

Table 1

Behavior-Count Data for Two Teachers

Data	Month 1	2	3	4	5	6
Teacher #21011						
Behavioral criticisms	7	3	3	19	16	26
Hours of observation	3	1	1	4	5	6
Empirical rate	2.3	3.0	3.0	4.7	3.2	4.33
Teacher #21052						
Behavioral criticisms	7	10	5	8	4	40
Hours of observation	1	3	2	4	6	5
Empirical rate	7.0	3.3	2.5	2.0	0.67	8.0

Note. The six months are November through April. Data were taken from the Texas Junior High study. See Appendix B for description of these data.

each month). A natural statistical model for the behavior-count data states that, for occasion i, the X_i are sampled from a Poisson distribution with rate parameter λ_i (see Appendix A). (Statistical procedures based on the Poisson distribution are applicable especially to the study of rare events, such as certain classroom-management behaviors.) The rates of occurrence of the target behavior (λ_1, . . ., λ_T) are the key behavioral parameters in the analysis of consistency over time for behavior-count data.

Bernoulli-trial Data

A novel and useful way to represent certain teacher behaviors is through the use of Bernoulli-trial data. Rather than recording only the frequency of a target behavior within a time interval, the investigator records the number of times the behavior occurs (the successes) among the opportunities for its occurrence (the trials).

Table 2 displays data adapted from Moon's (1969, 1971) study of elementary school science teachers (for a description of these data, see Appendix B). The target behavior is teacher questioning, in particular the asking of lower-order questions. For example on Occasion 5, 27 questions were asked by Teacher 8 of which 2 were lower-order questions, while on Occasion 6, 40 questions were asked of which 36 were lower-order questions.

In the Bernoulli-trial formulation, each question asked by the teacher constitutes a trial; the occurrence of a lower-order question is considered a success for that trial. The outcome of each Bernoulli trial (teacher question) is dichotomous; the outcome is one if the teacher question is a lower-order question and zero if it is not a lower-order question. The second row of the

Table 2

Bernoulli-trial Data for Two Teachers

Data	Occasion 3	4	5	6
Teacher No. 13				
Lower-order questions	8	11	12	13
All Teacher questions	24	49	35	75
Empirical proportion	.33	.22	.14	.26
Teacher No. 8				
Lower-order questions	23	6	2	36
All Teacher questions	66	29	27	40
Empirical proportion	.35	.21	.07	.90

Note. Data were taken from Moon's elementary-science study. See Appendix B for description of these data.

data for each of the two teachers in Table 2 contains the total number of trials (teacher questions) occurring on Occasion i, which is denoted by n_i. In the first row are the total number of successes (lower-order questions) on occasion i, denoted by X_i. (The X_i are the sum of the n_i dichotomous outcomes.)

For some teacher behaviors, use of the Bernoulli-trial representation is a useful alternative to the behavior-count representation. For example, on some days a teacher may ask more questions than on other days. Consequently, even if the teacher uses the same mix of higher-order and lower-order questions on each day, the number of lower-order questions would differ greatly from day to day, whereas the proportion of lower-order questions would remain relatively constant. The Bernoulli-trial representation may better reflect what teachers intend to do.

Furthermore, our formulation of teacher questioning behavior through the use of Bernoulli-trial data is consistent with the considerable empirical research on the degree to which teachers use higher-order versus lower-order questions. That is, the mix of questions is what some researchers consider to be educationally important, as opposed to only the frequency of particular types of questions. In recent reviews of research on teacher questioning (Winne, 1979; Redfield & Rousseau, 1981) interest centers on the effects of the contrast between teaching "dominated by fact questions" and teaching with "a greater proportion of higher cognitive questions" (Winne, 1979, p. 14). Earlier research on types of teacher questions is described in Gall (1970), Shavelson, Berliner, Ravitch & Loeding (1974), and Ryan (1973, 1974).

The Bernoulli-trial representation is also applicable to the analysis of contingent behaviors. For example, after a student has answered a question

correctly, the teacher may or may not praise that response. In representing this behavior sequence, teacher behavior would be dichotomized--the teacher either praises a student's correct response or does not praise that response. The correct student answer defines a trial; the occurrence of teacher praise in response to the student's answer is a success.

Another example of a behavior sequence for which the Bernoulli-trial representation is appropriate arises from the study of teachers' verbal behavior during reading lessons reported by Allington (1980). The behavior sequence of interest was the teacher's response (interruption or not) following an oral-reading error. In our formulation a student oral-reading error would define a trial for which the outcome is whether or not the teacher interrupts the student. Allington finds marked differences in teacher interruption behavior with students of high and low reading ability and advocates, among a number of directions for further research, that "research should identify whether teachers are consistent across time [on interruption behavior]" (p. 376).

Statistical models for Bernoulli-trial data are based on the binomial distribution (see Appendix A). For Occasion i, the probability of a success (e.g., a lower-order question) in a single trial is π_i. The $\pi_1, \ldots, \pi_T$ are the key behavioral parameters in the analysis of consistency over time for Bernoulli-trial data.

An extension of the Bernoulli-trial formulation, using the multinomial distribution, would be appropriate for a trial having more than two possible outcomes. For example, Moon (1969) actually classified teacher questions into five categories: recall facts, see relationships, make observations, hypothesize, test hypotheses. In Table 2 the five categories have been

collapsed into two categories: recall facts (lower-order questions) versus other kinds of questions. We emphasize the binomial model because of its simplicity and wide applicability. Use of the multinomial model may be advantageous when the multiple outcomes of a trial can be clearly categorized.

Quantitative Measures

The third type of data on teacher behavior are quantitative measures. Examples of quantitative measures include high-inference measures such as observer ratings of teacher behavior (perhaps averaged over raters), derived quantities like the indirectness-directness ratio of Flanders (1970), or a quantity such as the number of minutes of transition time between classroom activities. Tables 3 and 4 display examples of quantitative measures. In Table 3 are values of Flanders's indirectness-directness (I/D) ratio over five occasions for two of the teachers observed by Moon (1969). In Table 4 are monthly averages of ratings of positive affect (rated on a zero-to-four scale) in two English classes from the Texas Junior High School Study (Evertson & Veldman, 1981). (For a description of these data, see Appendix B.) These two classes were taught by the same teacher.

We model these quantitative measures by assuming that, on Occasion i, the measure X_i has a Gaussian (normal) distribution with mean μ_i and variance σ_i^2. (Sometimes a standard data transformation may be useful to make the assumption of a Gaussian distribution more reasonable.) The means at each occasion $\mu_1, \ldots, \mu_T$ are the key behavioral parameters in assessing consistency over time for quantitative measures.

Table 3

Indirectness-Directness (I/D) Ratios for Two Teachers

	Occasion				
Teacher	2	3	4	5	6
6	0.87	0.88	0.88	1.29	0.72
1	1.11	6.30	1.36	0.65	4.39

Note. Data were taken from Moon's elementary-school science study. See Appendix B for a description of these data.

Table 4

Ratings of Positive Affect for Two Classes

	Month					
Class	1	2	3	4	5	6
24033	3.00	3.33	2.57	2.00	1.33	1.00
24035	3.50	3.33	3.00	3.00	2.33	2.50

Note. Data were taken from the Texas Junior High School Study. See Appendix B for a description of these data.

Representing Stability for the Behavior of an Individual Teacher

The temporal stability (consistency over time) of the behavior of an individual teacher, which is the subject of *Question A*, requires that the behavior of the teacher remain unchanged over time. In the homogeneity hypotheses below, the consistency over time of the behavior of an individual teacher is formally represented in terms of the behavioral parameter for each type of data.

Behavior-count data. The rate parameter, λ_i, of the Poisson distribution is constant over occasions: $\lambda_i = \lambda$ for all i.

Bernoulli-trial data. The parameter, π_i, of the Binomial distribution is constant over occasions: $\pi_i = \pi$ for all i.

Quantitative measures. The parameter μ_i of the Gaussian distribution is constant over occasions: $\mu_i = \mu$ for all i.

Although these representations of stability are straightforward, possible departures from stability are numerous and complex. Figure 1 displays different ways in which a homogeneity hypothesis can be violated. For each of the four plots, the behavioral parameter (λ_i, π_i, or μ_i) is plotted on the vertical axis, and time (ordered occasions of observation) is plotted on the horizontal axis. The homogeneity hypothesis is satisfied in the upper-left quadrant of Figure 1; the label, "Absolute Invariance," is after Wohlwill (1973, Figure 12.2). In this quadrant the behavioral parameter is unchanging over occasions.

The other three quadrants of Figure 1 depict specific configurations in which the homogeneity hypothesis is not satisfied. These configurations are presented to stress that rejection of a homogeneity hypothesis can be due to different forms of heterogeneity. In the upper right quadrant the behavioral parameter follows a systematic time trend. (If this particular configuration were "detrended," the adjusted behavioral parameter would satisfy the

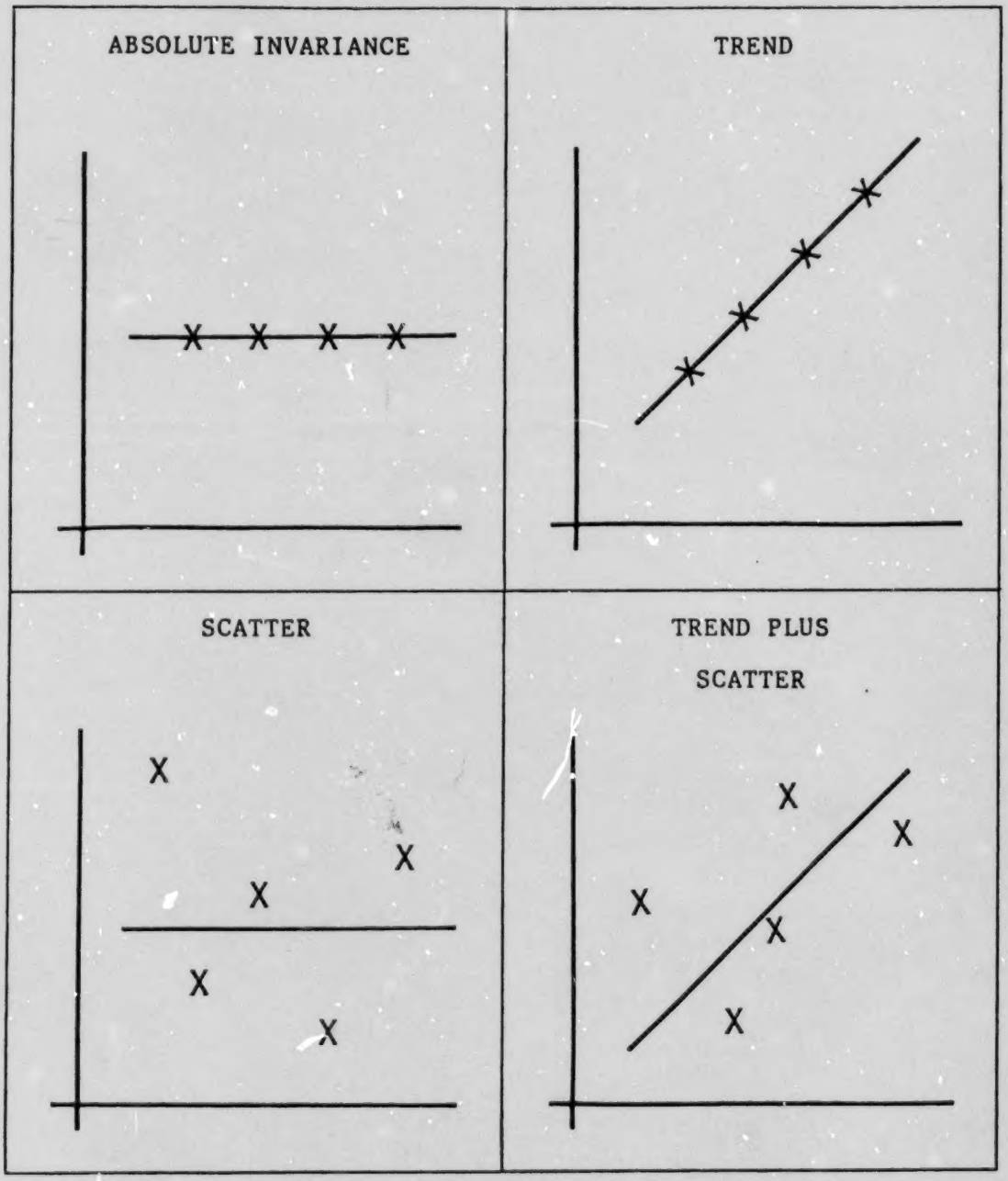


Figure 1. Illustrations of stability (consistency over time) and departures from stability; in each quadrant, the vertical axis is the behavioral parameter and the horizontal axis is time.

homogeneity hypothesis.) In the lower left quadrant, there is no time trend, but the scatter of the behavioral parameter violates the homogeneity hypothesis.[6] In the lower right quadrant, both the systematic time trend and the scatter contribute to the heterogeneity over time.

Statistical Procedures for the Behavior of an Individual Teacher

The statistical procedures for addressing *Question A*--the consistency over time of the behavior of an individual teacher--have two main functions: (a) to assess the viability of the relevant homogeneity hypothesis, and (b) to estimate the amount of heterogeneity. The exact form of the appropriate statistical procedure differs for the three types of data; each type of data is considered in turn.

Statistical Procedures for Behavior-Count Data

Testing homogeneity. One obvious way to assess the consistency over time of the behavior of an individual teacher is to perform, using that teacher's data, a test of the homogeneity hypothesis--$\lambda_i = \lambda$ for all i--against the general alternative of non-homogeneity, that not all the λ_i are equal. The traditional test statistic for this null hypothesis is presented in equation A1 of Appendix A. This test statistic assesses whether the observed counts (the X_i) are more spread out than would be expected under the homogeneity hypothesis, provided that the assumptions of the Poisson model hold.

---

[6] In statistical terms this quadrant represents a stationary, doubly-stochastic process, such as a doubly-stochastic Poisson process where the λ_i are themselves realizations of a stationary stochastic process (see for example, Cox & Lewis, 1966, Section 4.7).

Estimating heterogeneity. A useful supplement to testing the homogeneity hypothesis is to estimate the amount of heterogeneity in the teacher's behavior. This is represented by the variance of the λ_i, σ_λ^2. Because the homogeneity hypothesis posits that σ_λ^2 is zero, an estimate of σ_λ^2 is a natural measure of heterogeneity. This estimate of σ_λ^2 for a teacher can be considered a description of that teacher's behavior in the same way that the average rate of behavior describes that teacher. Cox (1955, Section 5.3) developed a variance component estimation procedure for σ_λ^2 that does not require choosing a specific distribution for the λ_i. Alternatively, when all the observation periods are the same length and the λ_i can be assumed to have a gamma distribution, estimation procedures based on the negative binomial distribution are appropriate (see Appendix A).

Trends. One important violation of the homogeneity hypothesis is a time-dependence for the rate of behaviors (as shown in the upper- and lower-right quadrants of Figure 1). For example, the frequency of call-outs by students may decline systematically over the course of a school year. Methods for modeling and analyzing the time dependence of the λ_i are presented in Cox and Lewis (1966, Chapter 3).

Examples. To illustrate statistical procedures for behavior-count data, analyses of the data in Table 1 are presented in Table 5. For each teacher the empirical rates of behavioral criticism, $\hat{\lambda}_i = X_i/b_i$, are shown in Table 1. Displayed in Table 5 are the test statistic values for the homogeneity hypothesis (from equation A1 of Appendix A) and also estimates of the mean and variance of the distribution of the λ_i.

Table 5

Heterogeneity-in-Counts Analysis
for the Data from Table 1

Teacher	Homogeneity test statistic[a]	Estimated moments of the λ distribution: Mean	Variance
21011	3.96 (5)	4.77	0.00
21052	49.0 (5)	3.92	6.95

[a]Numbers in parentheses are the degrees of freedom for the test statistic.

Although the two teachers have comparable average rates of occurrence for the target behavior (about four to five instances of behavioral criticism per hour), inspection of the $\hat{\lambda}_i$ in Table 1 shows that Teacher 21052 appears to be far less consistent over time than Teacher 21011. More formally, the test statistic for the homogeneity hypothesis falls far short of statistical significance for Teacher 21011 and is highly significant for Teacher 21052. The estimate of σ_λ^2 for Teacher 21011 is set to zero because the variance component estimate was negative. (A zero estimate is consistent with failure to reject the homogeneity hypothesis.) Teacher 21052 exhibits considerable heterogeneity, with an estimate for σ_λ^2 of 6.95.

In an analysis of a collection of teachers, we apply these statistical procedures to the data of each teacher separately. A major purpose of the analysis of a collection of teachers is to examine the consistency over time of individual teachers in relation to each other, that is, to investigate individual differences in consistency of behavior over time. For an

example of an analysis of a collection of teachers we used another low-inference variable from the Texas Junior High School Study, namely, product (lower-order) questions (i.e., "questions that have a specific correct answer which can be expressed in a single word or short phrase," Evertson & Veldman, 1981, p. 157). In Table 6 are results of a heterogeneity-in-counts analysis for 34 English teachers and, separately, for 22 mathematics teachers, using the behavior-count data for the product questions variable.

Table 6

Summary of Heterogeneity-in-Counts Analysis for Two Collections of Teachers

	English		Math	
Stem	Estimated mean of the λ distribution	Estimated variance of the λ distribution[a]	Estimated mean of the λ distribution	Estimated variance of the λ distribution[a]
30				
28		15,		
26		,36		
24	4,			
22	8,	9,	3,	
20	3,4		,7	,4
18				4,
16	,2	8,	6,7	
14	3,1116	,3	3,	
12	359,6	,7	03,0	3,
10	334,9	0,79	,68	0,
8	0,9	789,9	,4	1,
6	01277,	9,	06,	,17
4	2,4	1,167	05,456	,1
2	23,13	889,3		123,35
0	68,6	00133679,23	689,9	111355,057

Note. Data were taken from the Texas Junior High School Study. See Appendix B for description of these data.

[a]Multiply stem.leaf by 10.0 .

The stem-and-leaf diagrams (Tukey, 1977) in Table 6 show the empirical distributions of the estimates of μ_λ and σ^2_λ over each group of teachers. For each teacher two quantities are displayed, the estimate of the average (over occasions) hourly rate of product questions ($\hat{\mu}_\lambda$) and the estimate of the variability (over occasions) in the rate of product questions ($\hat{\sigma}^2_\lambda$). For the mathematics teachers the largest $\hat{\mu}_\lambda$ is 22.3 and the largest $\hat{\sigma}^2_\lambda$ is 214, whereas the smallest $\hat{\mu}_\lambda$ is 0.6, and the smallest $\hat{\sigma}^2_\lambda$ is 0.5 (rounded up to 1 in Table 6). Table 6 does not reveal, for example, that the English teacher (teacher 21023) with the largest mean rate of product questions, a $\hat{\mu}_\lambda$ of 24.4, also has the second largest estimated heterogeneity, a $\hat{\sigma}^2_\lambda$ of 281. The homogeneity hypothesis is rejected (at level .05) for all of the 22 mathematics teachers and for all but one of the 34 English teachers.

The stem-and-leaf diagrams for $\hat{\mu}_\lambda$ illustrate what has been noted occasionally in the literature--that teachers differ considerably in their *average rates of behavior*, for variables such as product questions. The stem-and-leaf diagrams for $\hat{\sigma}^2_\lambda$ reveal a new aspect of how teachers differ--teachers may also differ considerably from one another in the *consistency of their rates of behavior*.

Statistical Procedures for Bernoulli-Trial Data

Testing homogeneity. For Bernoulli-trial data, the appropriate procedure is to test the homogeneity (null) hypothesis--$\pi_i = \pi$ for all i--against the general alternative of non-homogeneity, that not all the π_i are equal. A traditional test statistic for this null hypothesis is the binomial "index of dispersion" (see equation A6 in Appendix A).

Estimating heterogeneity. The variance of the π_i distribution, σ_π^2, represents the heterogeneity in the behavior of an individual teacher over occasions. (The homogeneity hypothesis states that σ_π^2 is zero.) Various estimates of σ_π^2 have been developed in the statistical literature. In Appendix A, the estimates developed by Hendricks (1935), Kleinman (1973), and Robertson (1951) are described.

Trends. The existence of a time dependence, or trend, in the π_i is one important type of violation of the homogeneity hypothesis. A standard test for linear trend was developed by Armitage (1955).

Examples. To illustrate the statistical procedures for Bernoulli-trial data, analyses of the data in Table 2 are presented in Table 7. For each teacher the empirical proportions of lower-order questions, $p_i = X_i/n_i$, are given in Table 2. Displayed in Table 7 are a test statistic for the homogeneity hypothesis (from Expression A6) and estimates of the mean and variance of the

Table 7

Heterogeneity-in-Proportions Analysis for the Data from Table 2

Teacher	Homogeneity test statistic[a]	Estimated moments of the π distribution	
		Mean	Variance
13	5.06 (3)	.247	.001
8	58.1 (3)	.384	.095

[a]Numbers in parentheses are the degrees of freedom for the test statistic.

distribution of the π_i (from Kleinman, 1973). Teacher 8 shows appreciable heterogeneity over the four occasions, whereas the data for Teacher 13 are consistent with the homogeneity hypothesis and the estimate of σ_π^2 is nearly zero.

Tables 8 and 9 present analyses of lower-order questions for two additional teachers observed by Moon. Clearly, neither teacher's behavior is consistent

Table 8

Bernoulli-trial Data
Lower-order Questions

Data	Occasion 3	4	5	6
Teacher #15				
Lower-order questions	5	7	15	16
All Teacher questions	56	33	45	28
Empirical proportions	.09	.21	.33	.57
Teacher #11				
Lower-order questions	17	8	7	17
All Teacher questions	34	24	19	93
Empirical proportions	.50	.33	.37	.18

Note. Data were taken from Moon's elementary-school science study. See Appendix B for description of these data.

over time; each teacher shows a statistically significant time trend over the four occasions of observation. Teacher 15 shows a positive trend over time, and teacher 11 shows a negative trend. For estimation of these trends, the four occasions were treated as equally spaced in time; use of the exact dates of observation yields similar results. These opposite, significant trends reinforce our point (see Figure 1) that a violation of absolute invariance does not necessarily imply haphazard fluctuation.

Table 9

Trend and Heterogeneity-in-Proportions Analysis for Two Teachers

Teacher	Homogeneity test statistic[a]	Estimated moments of the π distribution		Analysis of linear trend	
		Mean	Variance	Est. slope	Std. error
15	23.9 (3)	.296	.0265	.15	.031
11	13.3 (3)	.336	.0076	-.10	.028

[a]Numbers in parentheses are the degrees of freedom for the test statistic.

Table 10 presents analyses of Bernoulli-trial data for two collections of teachers from investigations by Moon and by Trinchero (see Appendix B). A question asked by the teacher constitutes a trial and a lower-order question counts as a success. Considerable individual differences in consistency exist in both collections of teachers. The estimates of σ^2_π range between 0.0001 and 0.0947 for the Moon data and between 0.0 and 0.0824 for the Trinchero data. Also, Table 10 reveals that the estimates of μ_π for the teachers in the Moon

study differ considerably from the estimates of μ_π for teachers in the Trinchero study. The homogeneity hypothesis will not be rejected for teachers whose estimate of σ^2_π is very close to zero. For the Moon data, the homogeneity hypothesis is rejected (at level .05) for 13 of the 16 teachers. For the Trinchero data, the homogeneity hypothesis is rejected (at level .05) for only 5 of the 22 teachers.

Table 10

Summary of Heterogeneity-in-Proportions Analysis for Two Collections of Teachers

	Moon		Trinchero	
Stem	Estimated mean of the π distribution[a]	Estimated variance of the π distribution[b]	Estimated mean of the π distribution[a]	Estimated variance of the π distribution[b]
10				
9		5	011	
8			11225799	2
7		0	023577	
6			2668	13
5			4	
4	49	9		
3	0112468	2		
2	257779	177		5
1	5	22446		37
0		1388		000000000000012

[a]Multiply stem.leaf by 0.1 .
[b]Multiply stem.leaf by 0.01 .

The Trinchero data provide an excellent example of the consequences of considering the target behavior (lower-order questions) as Bernoulli-trial data rather than as behavior-count data. In contrast to the analysis of the Bernoulli-trial data (for which the homogeneity hypothesis is rejected for 5 of

22 teachers), analysis of only the behavior-counts for lower-order questions leads to rejection (at level .05) of the homogeneity hypothesis for 19 of the 22 student teachers! These Bernoulli-trial results provide empirical support for the contention that teachers may be consistent in their questioning behavior, a finding that would not be obtained if the behavior-count representation were used.

Statistical Procedures for Quantitative Measures

Testing homogeneity. A direct test of the homogeneity hypothesis for quantitative measures (all the μ_i are equal) is more difficult to obtain than the statistical tests for the homogeneity hypotheses with behavior-count data or Bernoulli-trial data. Recall that we adopted a model for quantitative measures which states that X_i is drawn from a Gaussian distribution with mean μ_i and variance σ_i^2. Because in the Gaussian distribution the mean and variance are unrelated, a determination of whether the X_i are more spread out than expected (under the homogeneity hypothesis) is impossible without additional information on the σ_i^2. (That is, for the Poisson model the mean of X_i is $b_i\lambda_i$, which is also the variance, and for the binomial model the mean is $n_i\pi_i$ and the variance is $n_i\pi_i(1 - \pi_i)$.)

Consider that the μ_i have a distribution with mean θ and variance κ^2. The homogeneity hypothesis states that $\kappa^2 = 0$. If the distribution of the μ_i is Gaussian, then the compound distribution of the X_i is Gaussian (see Johnson & Kotz, 1970, Section 13.7.2) with mean θ and variance $\sigma^2 + \kappa^2$ (setting $\sigma_i^2 = \sigma^2$). Thus, if σ^2 were known, then a statistical test of the homogeneity hypothesis could be conducted by testing the hypothesis that $\sigma_X^2 \leq \sigma^2$ against the alternative that $\sigma_X^2 > \sigma^2$ (i.e., $\kappa^2 > 0$) using standard chi-square methods. In

some cases the most useful assumption may be that $\sigma_i^2 = 0$ for all i.

Indices of stability. A natural measure of heterogeneity would be an estimate of κ^2. A number of alternative quantities may be useful in situations where an estimate of κ^2 is not available. Certainly, the sample variance or standard deviation of the X_i for each teacher permits some comparison of the consistency over time of the behavior of different teachers. For example, in her analysis of children's cognitive development, Bayley (1949) computed the standard deviation of each child's IQ score over multiple testings. This standard deviation was termed a "Lability Score." In addition, children were characterized as "labile" or "stable" if they were in the upper or lower quartiles, respectively, of the empirical distribution over all children of these standard deviations. Similarly, Flanders (1969) used the standard deviation of the indirectness-directness (i/d) ratio across situations and occasions to obtain an "index of flexibility" for each of 20 teachers.

Measures related to the standard deviation such as the coefficient of variation, Gini's mean difference and the coefficient of concentration (see Kendall & Stuart, 1969) provide similar descriptive information on the consistency over time of each teacher. Alternative indices can be adapted from the early statistical studies of stability of a statistical series (see Forsyth, 1932, 1937; Bortkiewicz, 1931), which are derived from Lexis theory, an active topic in late-nineteenth-century statistical work.

Trends. The simple X on t regression function for an individual teacher can be used to detect time trends for quantitative measures. The product-moment correlation r_{Xt} indicates the strength of the linear time trend for the quantitative measure. With many observations on each teacher, more complex time

dependencies can be investigated by fitting a polynomial or a non-linear function of time.

Examples. The data in Tables 3 and 4 are used to provide examples of the examination of consistency over time for individual teachers using descriptive statistics for the quantitative measures. Teacher 6 from Table 3 has a standard deviation for the indirectness-directness (I/D) ratio of .21, whereas Teacher 1 has a standard deviation of 2.46. For the 16 teachers in Moon's study, the standard deviation of the indirectness-directness (I/D) ratio ranges between .15 and 2.46, with a median of .74. Relative to the other teachers observed in this study, Teacher 6 appears consistent over time, whereas Teacher 1 has by far the least consistent behavior. Neither of these teachers show a time trend for the indirectness-directness (I/D) ratio; the magnitude of r_{Xt} is less than .10 for both teachers.

In Table 4 are ratings of positive affect from two classes taught by the same junior-high-school English teacher. Notable in these data is that r_{Xt} is -.96 and -.94 for the two classes respectively--the strongest associations seen in the 76 English classes in the Texas study. Naturally, both this teacher's classes show highly negative and statistically significant time trends.

Consistency of Individual Differences

Question B, concerning the consistency or maintenance of individual differences over time, has been pre-eminent in empirical studies of stability. One statement of interest in *Question B* is seen in Shavelson and Dempsey-Atwood (1976): "Although it is possible to consistently rank order teacher performance at one point in time it is an empirical question as to whether this rank ordering is stable" (p. 554). Our purposes in this section are to state

explicitly what is meant by the consistency of individual differences over time and to present statistical procedures for assessing the degree of consistency of individual differences.

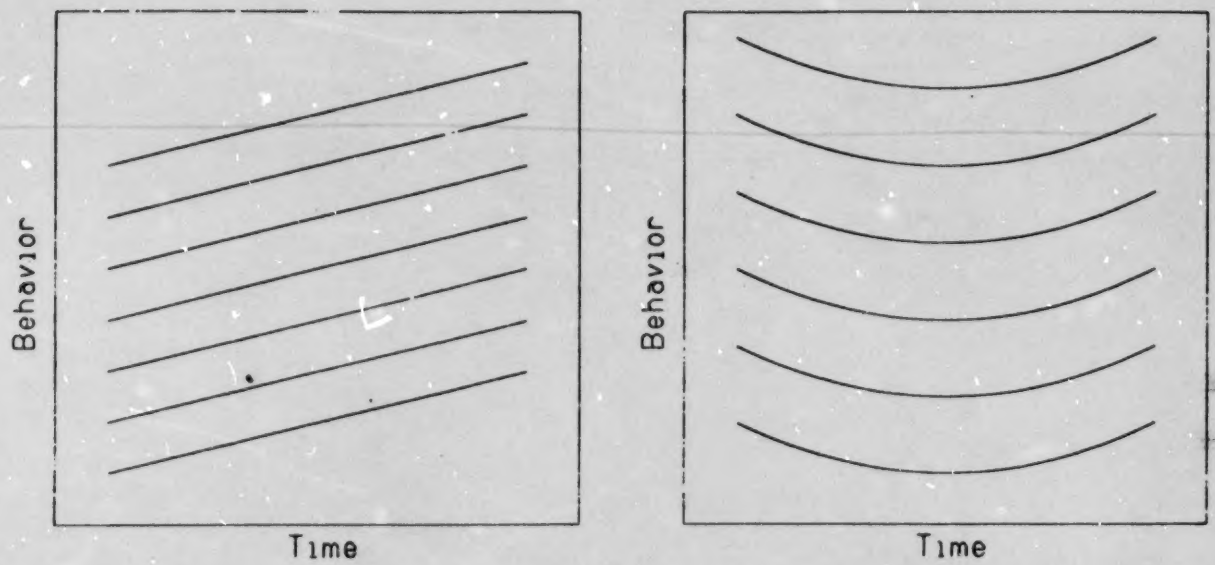
Representing Consistency of Individual Differences

The consistency of individual differences over time is a property of the collection of individual time paths for the target behavior. The individual time path is a representation of each teacher's behavior as a systematic function over time. (For convenience, this exposition will focus on functions of quantitative measures over time.) Note that a time dependence of the behavior of an individual teacher constitutes a violation of the relevant homogeneity hypothesis (see Figure 1).

Figure 2 displays two examples of (perfectly) consistent individual differences over time. In Figure 2a the time path of the target behavior for each teacher is a straight line. Wohlwill (1973, Figure 12-7) uses a similar diagram to represent the "Preservation of individual differences." In Figure 2b individual differences are maintained in a collection of curvilinear time paths.

Three criteria can be used to define perfectly consistent individual differences over time: (a) absolute vertical distance between pairs of time paths unchanged over time (i.e., all time paths parallel), (b) relative distance between time paths unchanged over time (i.e., percentile rank of points on each time path constant over time), and (c) rank order of time paths maintained over time (i.e., time paths do not intersect). Criterion (a) is more strict than criterion (b), which is, in turn, more strict than criterion (c). Both frames in Figure 2 show consistency under all three criteria.

The biometric literature provides methods that can be applied to the study



(2a) (2b)

Figure 2. Two illustrations of (perfect) consistency of individual differences over time.

of individual differences among teachers. In this literature, the consistency of individual differences over time is described by the term "tracking." A major substantive concern in medical research is whether blood pressure tracks--whether children with relatively high blood pressure compared to a representative group of children maintain that position over time and become high-risk adults. The empirical question of whether individual differences are preserved over time is precisely the question that has been dominant in research on teaching.

Statistical Procedures--Indices of Tracking

Perfect tracking of teacher behavior is probably rare, even under the least restrictive criterion that rank order be preserved over time. The degree of consistency of individual differences over time can be described by an index of tracking. Use of an index of tracking provides important advantages over the correlational analyses common in research on teaching. First, an index of tracking allows assessment of the consistency of individual differences over more than two time points. Second, an index of tracking incorporates explicit statistical models for the individual time paths and thus is applicable when time trends in behavior are present.

Foulkes-Davis γ. Foulkes and Davis (1981) propose an index of tracking, denoted by γ, which reflects "the maintenance over time of relative ranking within the response distribution" (p. 439). Thus perfect tracking occurs when a collection of individual time paths do not intersect in a specified time interval. The index of tracking is defined as the probability that two randomly chosen time paths do not intersect during a specified time interval. No tracking is said to occur if $\gamma < .50$. To estimate γ, polynomial time trends are

fitted to the data for each individual; the estimate of γ is then determined from the pairwise intersections of these fitted time trends.[7] Note that, when measurements are available at only two points in time, $\hat{\gamma}$ reduces to an estimate of Kendall's probability of concordance; consequently, $\hat{\gamma}$ may be thought of as a generalization of a rank correlation coefficient (Foulkes & Davis, 1981, Section 2).

McMahan τ. An alternative index of tracking appears in McMahan (1981). McMahan defines tracking as follows: "For each individual, the expected value of the relative deviation from the population mean remains unchanged over time" (p. 449), a definition closely related to the maintenance of percentile rank over time. The index of tracking represents the degree to which this definition is satisfied by the data. Specifically, τ represents the variance in X (corrected for within-subject error) explained by the individual's relative deviation from the population mean.[8] Unfortunately, in behavioral data, the correction for within-subject error may often overstate the real measurement error variance, making τ less attractive than γ for analyses of teacher behavior. That is, the correction for errors of measurement in the estimation of τ is model dependent, and unless the correct model for the time trend in the target behavior is fitted, this index of tracking is likely to be seriously inflated. Comparisons of τ and γ for various data sets are provided by Rogosa

---

[7]The estimate of the index of tracking is model dependent in the sense that different degrees of tracking will be seen when different functional forms for the individual time paths are fitted (e.g., quadratic vs. cubic fits).

[8]For measurements at only two points in time, the index τ is a product-moment correlation, corrected for attenuation. In the special case in which the variance of the measure is the same at each of multiple points in time, the index is simply the average of the pairwise, disattenuated correlation coefficients.

and Willett (1983).

Example. Using data on product (lower-order) questions for each of six months during the school year the consistency of individual differences for teachers in the Texas Junior High School Study (Evertson & Veldman, 1981) was assessed. First, the observed rate of product questions on each occasion was transformed to $\sqrt{\text{rate} + 3/8}$. (This transformation is an effective normalizing and variance-stabilizing transformation for Poisson variates.) Then, for each teacher a straight-line time trend was fitted to the transformed rate of product questions. The collection of these fits for the 25 English teachers is displayed in Figure 3. Although these fits have a considerable number of intersections, the figure does show that the fitted trends for some teachers remain consistently high, while the trends for others are consistently low. For these data, the estimate of Foulkes-Davis γ is .71 with an approximate standard error of .02; this estimate of γ reflects reasonably strong consistency of individual differences among teachers for product questions.

Reconsidering Previous Approaches and Methods

In the research-on-teaching literature three approaches to assessing stability have dominated work on stability of teacher behavior: (a) computation of correlations among observation times (e.g., Brophy, Coulter, Crawford, Evertson, & King, 1975), (b) application of generalizability theory (e.g., Erlich & Shavelson, 1978) and (c) estimation of occasion effects in repeated-measures analysis of variance (e.g., Evertson & Veldman, 1981). None of these approaches explicitly addresses *Question A*. Furthermore, the information that these approaches provide on *Question B* is limited and sometimes misleading.

Each of the three approaches uses the same basic data. We denote these data by X_{ij}, an individual datum being the measurement obtained for Teacher j

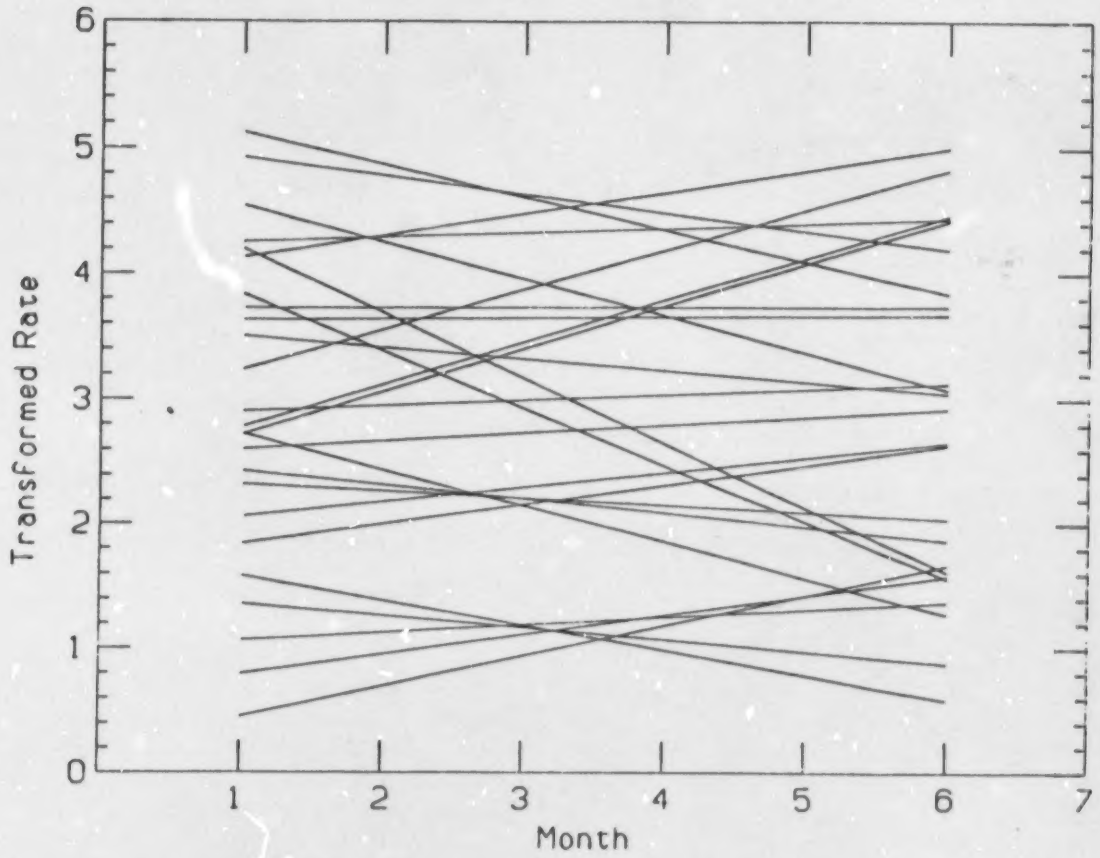


Figure 3. A collection of straight-line time trends in rate of product questions for 25 English teachers from the Texas Junior High School Study.

(j=1, . . . ,n) on Occasion i (i = 1, . . . ,T). The different approaches employ different statistical models for the X_{ij}, and consequently in each approach different summaries of the X_{ij} are sought. Below we describe each approach and give examples from empirical research. We are particularly interested in what can or cannot be learned from the statistical methods used in these approaches. Often a considerable gap, or even a contradiction, exists between the stated research question and the statistical methods employed to assess stability.

Correlations Among Observation Times

The correlational approach, by far the most common approach to assessing stability of behavior, is seen as an extension of test-retest reliability. Most often, studies using correlations over observation occasions to assess temporal stability use only two occasions of observation and report time-one, time-two correlations as measures of stability. Some studies having observations on more than two occasions have adapted the test-retest correlation approach to multiple time points. Common to all these studies is the willingness of investigators to correlate most anything; whether the data be behavior-counts, proportions, rates, complex derived indices (such as ratios), or high-inference ratings, the product-moment correlation is used with little concern for the distributional assumptions in statistical inference,[9] for the adequacy of a measure of linear

[9]Although the use of the product-moment correlation as a descriptive statistic does not explicitly depend on any assumptions about the bivariate distribution of the time-one, time-two data, researchers often report, in addition to the correlation coefficient, the results of statistical inference procedures (e.g., p-values for the null hypothesis of zero correlation) that do depend crucially on the validity of the underlying assumption of a bivariate normal distribution. In addition, it seems appropriate to remark that rejection of the null hypothesis of a zero time-one, time-two correlation (say, at level .05) is very weak evidence of stability. However, such a criterion, used formally or informally, is widespread in the literature.

association, or for the usefulness of data transformations.

Time-one, time-two correlations. The time-one, time-two correlation, or test-retest correlation, is the measure of stability of teacher behavior used in almost all empirical research. The data consist of measurements on the target behavior on two occasions for each of n teachers. The stability measure is the usual product-moment correlation between the measurements at time one and the measurements on the same teachers for time two, written as $r_{X_1X_2}$.

In their influential chapter, Medley and Mitzel (1963) define the "stability coefficient" to be "a correlation between scores based on observations made by the same observer at different times" (p. 254). They add that "the coefficient of stability tells us something about the consistency of the behavior from time to time" (p.254). To amend Medley and Mitzel's interpretation of the "stability coefficient," $r_{X_1X_2}$ *only* tells us something about the consistency of *individual differences* in behavior from time to time. That is, the correlation provides information only on *Question B*.

Applications. A typical example of the correlational approach is seen in Brophy, Coulter, Crawford, Evertson, and King (1975). They report analyses of data obtained by observing 19 second- and third-grade teachers using the Classroom Observation Scales. Teachers were observed for two mornings and two afternoons in the first year of the study and on 14 occasions in the second year of the study. The observations for multiple time periods were reduced to two measurements by computing each teacher's mean score for each year. For a number of target behaviors, correlations were computed between these mean scores over the two years.

More extensive examples can be found in the Shavelson and Dempsey-Atwood

(1976) compendium of time-one, time-two correlations. Among the teaching behaviors included in the Shavelson and Dempsey-Atwood review are various kinds of teacher questions for which they conclude: "the data from seven studies indicate that teacher questioning behavior is unstable" (1976, p. 605). Among the many studies reviewed are those by Moon and Trinchero, the data from which have been used in this paper. Interestingly, Shavelson and Dempsey-Atwood (1976, Table 1), using two occasions from the Trinchero data, report "stability coefficients" of 0.17 for both lower-order and higher-order teacher questions. Another example of the use of time-one, time-two correlations is provided by the data on "conventional teachers" in Moon's study (1969, 1971) (see Appendix B) for which Shavelson and Dempsey-Atwood report "stability coefficients" of 0.42 for lower-order (factual) questions and 0.64 for the indirectness-directness (I/D) ratio.

Multiple occasions. Similar correlational strategies for addressing *Question B* have sometimes been adopted when data from more than two observation occasions are available. The assumption that the test-retest correlation is the same for all pairs of time points is the basis for the use of an average correlation over all pairs of time points as a "stability coefficient." The multiple occasions are employed merely to replicate the test-retest correlation, and the $T(T-1)/2$ correlations among all time points (the $r_{X_i X_{i'}}$) are averaged arithmetically or, more appropriately, by using Fisher's z-transformation. Alternatively, an intraclass correlation coefficient, based on the model stated in Ebel (1951), can be used in place of the average correlation.

Most importantly, the time ordering of the observations is ignored in this averaging of correlations to form a stability coefficient, and thus time trends

in the behavior cannot be accommodated. With multiple time points, if time trends in the target behavior exist, then the indices of tracking are far superior to test-retest correlations for addressing *Question B*.

An example of the correlational approach with data from multiple occasions is the reanalysis of the teacher questioning data for the final four occasions from Moon's study of elementary science teachers which is presented in Rosenshine (1973, Table 2). He reports both an average correlation (using Fisher's z-transformation) and an intraclass correlation for each of the five types of teacher questions. As each of the five 4 x 4 correlation matrices include both positive and negative elements of moderate magnitude, the intraclass correlation yields "stability coefficients" of zero (or even negative coefficients if the average correlation is used). Other examples of this approach to the analysis of multiple-time-point data are found in Shavelson and Dempsey-Atwood (1976).

Generalizability Theory

Generalizability theory provides a second approach to the assessment of stability of teacher behavior. The data for generalizability analyses consist of two or more observations of a target behavior on each of n teachers, the observations being made by κ raters. For simplicity, we will consider only the case of κ=1 rater, and thus the data are the X_{ij} as in the previous section. (The assumption $\kappa = 1$ is an extreme simplification of generalizability-theory methods, but this restriction is useful for our exposition of applications to stability.) The measure of stability is the coefficient of generalizability.

Generalizability theory is an extension of classical test theory which features the separate estimation of several sources of variation in the

observations. Applications of generalizability theory to the stability of teacher behavior have been advocated by Erlich and Borich (1979), McGaw, Wardrop and Bunda (1972), and Shavelson and associates (Erlich & Shavelson, 1978; Shavelson & Dempsey-Atwood, 1976; Shavelson & Webb, 1981). A useful interpretation of the generalizability coefficient is that it indicates the ability to detect differences among teachers' average behavior, that is, the generalizability coefficient answers the question, "How well can the average behavior of each teacher be located relative to other teachers' average behavior?" As discussed earlier in this paper, the ability to differentiate teachers on their (average) behavior is crucial for the success of process-product research.

A major limitation of generalizability theory for assessing temporal stability is that generalizability theory ignores time trends in behavior by focusing on the time average of each teacher, the $\overline{X}_{\cdot j}$. (All variation of the X_{ij} about $\overline{X}_{\cdot j}$ is construed as "error.") Specifically, generalizability theory assumes a steady state for the behavior over time of each individual teacher: "Because our model treats conditions within a facet as unordered, it will not deal adequately with the stability of scores that are subject to trends" (Cronbach, Gleser, Nanda, & Rajaratnam, 1972, p. 364). A consequence of treating the time facet as unordered is that it must be assumed that all individuals are consistent over time, that is, *Question A* is answered affirmatively for each individual (see also Ebel, 1951, p. 409).

An analysis of variance model for generalizability-theory analysis (with one rater) is as follows:

$$X_{ij} = \mu + \beta_j + \epsilon_{ij} \qquad (i = 1,\ldots,T;\ j = i,\ldots,n).$$

where β_j represents the teacher effect. In terms of this one-way, random-

effects model the question addressed by the use of generalizability theory can be expressed precisely as, "Is σ^2_β (the variance component for teachers) big compared to the error?" The coefficient of generalizability for this model, $\sigma^2_\beta/(\sigma^2_\beta + \sigma^2_\varepsilon/T)$, measures the ability to distinguish among teachers' mean behavior over hypothetical replications of the measurement. The proper interpretation of this coefficient in the assessment of stability of behavior is that the coefficient of generalizability addresses *Question B* in a peculiar, conditional manner. Specifically, conditional on the steady state assumption, the coefficient of generalizability assesses individual differences among teachers. Consequently, the assumption that all individuals are consistent over time is crucial to the interpretation of the generalizability coefficient as a measure of stability of individual differences.[10]

In some applications of generalizability theory to analysis of teacher behaviors, a two-way analysis of variance model has been used:

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad (i = 1,\ldots,T;\ j = i,\ldots,n) .$$

The inclusion of the occasion effect, α_i, in this model is contrary to the statement of the steady state hypothesis. Although the occasion effect allows for trends in individual behavior, this model constrains all individual teachers to have the same time trend. The inclusion of the occasion effect in the model

---

[10]The assumptions underlying the use of the generalizability coefficient can be seen from the formulation of Ebel (1951) in which the observations "may be considered to consist of a true component and an error. The true component is constant in all [T] estimates for any one person but varies from person to person" (p. 409). That is, in Ebel's model which underlies the use of the intraclass correlation, all individuals are assumed to be consistent over time (on true score). The generalizability coefficient (with $\kappa = 1$) is simply Ebel's "reliability of average rating," which can be obtained by applying the Spearman-Brown formula to the intra-class correlation (see also Haggard, 1958, pp. 89, 134). (To link Ebel's formulation with the present discussion, occasions assume the role of raters.)

serves to reduce the error variance by removing the average trend over all teachers from the deviations of X_{ij} about $\bar{X}_{.j}$ for each teacher.[11]

Applications of generalizability theory to the analysis of classroom observation data can be found in Cronbach et al. (1972, Chap. 7); among their examples is a reanalysis of the data from Medley and Mitzel (1963), on 24 teachers for five occasions with two raters (see Table 7.1, p. 191). An additional application of generalizability theory to the analysis of teacher behavior is Erlich and Shavelson (1978) in which observations on five teachers in both reading and math lessons on three occasions (from a sub-study of BTES, Phase II, Sandoval, no date) were reanalyzed. Also, Erlich and Borich (1979), using data from five occasions on second- and third-grade teachers and the two-way analysis of variance model stated above, found that only 35 of the 167 classroom variables from the Teacher-Child Dyadic Interaction System were generalizable--the criterion being that "a generalizable variable was defined in this study as one for which a coefficient of generalizability of at least 0.7 could be obtained by observing the teacher on 10 or fewer occasions" (Erlich & Borich, 1979, p. 13). In contrast, Lomax (1982), using observations of student behavior during reading instruction from 11 elementary-level learning disability classrooms, obtained an "average stability coefficient" of .975 for 30 hour-long

---

[11] The distinction between the two analysis-of-variance models for the X_{ij}, in terms of their consequences for the generalizability coefficient, is identical to the considerations in Ebel (1951, p. 411) in his discussion of "removing between-raters [occasions] variance from the error term" (see Ebel, pp. 410-411, and Haggard, 1958, p. 89, for additional discussion). In some presentations of generalizability theory, this distinction is couched in the terminology of "relative decisions" versus "absolute decisions" (e.g., Shavelson & Webb, 1981, Section 1), and the generalizability coefficients for the one-way and two-way analysis of variance models are given in Equations 10 and 9, respectively, of Shavelson and Webb (1981).

observation periods.

Mean Teacher Time Path

A third approach to assessing stability of teacher behavior is the use of repeated-measures analysis of variance to investigate time trends in the mean over all teachers. The indication of stability in this approach is the lack of an occasion effect in repeated-measures analysis of variance. Thus, the question being addressed can be expressed as, "Is the mean behavior across teachers absolutely invariant over time?"

A way to relate this approach to the questions about stability is to consider an assumption that the time trend in the $\bar{X}_{i.}$ accurately represents the behavior of a typical teacher over time. Invoking this assumption, the absence of an occasion effect can be considered an affirmative answer to *Question A* for every teacher studied.

The data used in this approach to assessing stability are two or more observations on each teacher. The repeated-measures analysis of variance model for a single group of teachers can be written:

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (i = 1,\ldots,T;\ j = i,\ldots,n) .$$

The existence of an occasion effect, evidenced by a rejection of the null hypothesis that all α_i are equal, would indicate lack of stability. Although the model for the X_{ij} is the same as has been used in applications of generalizability theory, interest centers on a different parameter. The concern here is with the α_i; whereas in generalizability theory the spread of the β_j indicates the stability of the behavior.

An application of repeated-measures analysis of variance to assessing stability of teacher behavior is reported by Evertson and Veldman (1981) in

their analyses of teacher-behavior data from mathematics and English classes in the Texas Junior High School Study. (We use these data extensively in this paper.) The data analyzed by Evertson and Veldman are observations of behavior on six occasions, where the data for each occasion is an average of behavior observed within a month. Another example of a repeated-measures analysis of variance is that of Good, Cooper, and Blakey (1980), who examined teacher-student interaction over time.

Comparisons and Contradictions

Each of these three approaches to assessing stability of teacher behavior incorporates a different definition (rarely explicit) of stability. These differences in definition can be seen by considering the questions addressed by each approach. Test-retest correlations address the question, "Are individual differences among teachers maintained from Time 1 to Time 2?" Applications of generalizability theory address the question, "How well can the average behavior of each teacher be located relative to other teachers' average behavior?" Repeated-measures analysis of variance addresses the question, "Is the mean behavior across teachers absolutely invariant over time?" Though all three techniques have been used to assess "stability," the definition of stability (and thus the quantity being estimated) differs markedly among these three approaches.

Therefore, that assessments of stability from these three approaches may contradict each other is not surprising. For example, data on a target behavior that exhibits stability by virtue of a flat mean teacher time trend (i.e., no occasion effect in the repeated-measures analysis of variance) may have small

(or even negative) test-retest correlations.[12] Moreover, data on a teacher behavior may show a flat mean teacher time trend, and high test-retest correlation, yet show a small or negligible generalizability coefficient. Many other contradictions are possible.

Assessments of stability that follow from explicit statements of consistency over time are to be preferred to the less explicitly formulated procedures commonly used in research on teaching. Therefore, we recommend that statistical procedures based on the homogeneity hypotheses (especially estimates of heterogeneity) be used to address *Question A*, and that an index of tracking (in particular, Foulkes-Davis γ) be used to address *Question B*.

Stability Across Contexts

The stability of teacher behaviors across different contexts (e.g., subject matter or class composition) has also been of major research interest (e.g., Brophy et al., 1975; Shavelson & Dempsey-Atwood, 1976; Evertson, Anderson, Edgar, Minter, & Brophy, 1977). Research questions and statistical procedures resemble those for stability of teacher behaviors over time. Specifically, the two research questions for stability across contexts are

Question A*. Is the behavior of an individual teacher consistent across contexts?

---

[12] This contradiction is exhibited by the data on lower-order (recall facts) questions for Moon's SCIS teachers. The reanalysis of these data by Rosenshine (1973, Table 2) produced, for the final four occasions in Moon's study, a negative average correlation (-.18) and an intraclass correlation of 0.0 (with elements of the between-occasion correlation matrix ranging between -.49 and .45). Thus, the correlational approach indicates "low stability of individual teacher behavior across observations" (Rosenshine, 1973, p. 225). In contrast, a two-way analysis of variance of these data (with teachers and occasions as the factors) carried out by the authors yields an F-statistic of 1.48 (3 and 45 degrees of freedom) for the occasion effect. Thus by the mean teacher time path approach, the teacher behavior is found to be stable.

*Question B**. Are individual differences among teachers consistent across contexts?

As it has with stability of behavior over time, empirical research has focused exclusively on consistency of individual differences.

*Question A**. Statistical procedures for assessing the consistency of an individual teacher across contexts can be developed to test a null hypothesis of consistency. For example, with behavior-count data, a null hypothesis of consistency across two contexts states that the rate of the target behavior for teacher j is the same in context 1 as in context 2 (i.e., $\lambda_{1j} = \lambda_{2j}$). Or, for Bernoulli-trial data the null hypothesis of consistency across the two contexts for teacher j would be $\pi_{1j} = \pi_{2j}$.

Often, there may be multiple observations on the target behavior for each teacher in each context. For example, in the Texas Junior High School Study there are six observations obtained in each of two different class sections (the two contexts) for every teacher. For such data, relevant statistical procedures for testing the (null) hypothesis of consistency across contexts for each teacher can be found in Detre and White (1970) for behavior-count data, and in Kleinman (1973) for Bernoulli-trial data.

*Question B**. The consistency of individual differences between two contexts can be assessed by using a measure of association, such as the product-moment correlation coefficient. The measure of association summarizes the degree to which teachers who are high (in relation to the other teachers) on the target behavior in one context are also high on that target behavior in another context, and so forth. Perfect consistency of individual differences would be indicated by a correlation of 1.0 . Unlike occasions of observation, different

contexts seldom have an obvious ordering. (One example of ordered contexts might be low-, middle- and high-ability groups of students.) Hence consistency across contexts cannot usually be formulated using trends for each teacher.

Example. Data from the Texas Junior High School Study (Evertson & Veldman, 1981) include observations on teachers for two different classes during a school year (e.g., two different sections of eighth-grade English). These data can be used to address both *Question A** and *Question B** for differences in class composition. We illustrate the statistical methods for assessing stability across contexts by analyzing the behavior-count data for two commonly studied classroom variables: product (lower-order) questions and call-outs by students. Six months of data on two English classes for each of 25 teachers were used. For product questions, the null hypothesis that the teacher had equal rates of behavior in each of the two classes was rejected (at level .05) for 10 of the 25 teachers using Detre and White's (1970) test statistic. However, the consistency of individual differences among teachers was high, with a correlation of .84 between rates of product questions for the two classes. A different picture is seen for call-outs, where the hypothesis of consistency of individual teachers across contexts was rejected (at level .05) for only 4 of the 25 teachers. In this case a correlation of .47 across the contexts[13] does not indicate high consistency of individual differences among teachers.

These examples show that with stability across contexts, as with stability

---

[13]The reported correlations across contexts for both rate of product questions and callouts are actually correlations between transformed quantities, namely $\sqrt{\text{rate} + 3/8}$. The correlations using the raw--untransformed--rates were .75 and .39, as opposed to .84 and .47, respectively. Clearly, this transformation of the rate of behavior improves the linear association across contexts.

over time, it is important to specify an explicit research question and to apply a statistical procedure that corresponds to that research question. A correlational analysis of stability for product questions would have indicated high stability across contexts. Yet two-fifths of the individual teachers were not consistent across contexts.

Notes on Design

In the preceding sections, a number of statistical analyses for assessing stability have been presented. However, a major aspect of any study of stability, the design of the study, has not been explicitly discussed. Design considerations include a wide range of investigator decisions about how to carry out the study. In this section, we comment on three important design considerations: observation schedules, observation instruments, and homogeneous classroom contexts.

Observation Schedules

In designing an observation schedule, the investigator must determine how often and for how long the target behavior should be observed. Statistical considerations can be useful in determining the number of observation occasions and the length of the observation period on each occasion. Of course, normal classroom activities limit the possible length of any observation period (e.g., mathematics instruction cannot be observed for four hours on each occasion). Even so, a variety of observation schedules are possible, some of which will be more efficient than others.

Consistency of an individual over time. The statistical design problem is to devise an observation schedule that provides (within practical constraints)

as much information as possible about the parameter of interest (in particular, σ^2_λ or σ^2_π). For example, in addressing *Question A* with behavior-count data, the parameter σ^2_λ is of interest, both for testing the homogeneity hypothesis and in estimating heterogeneity. For testing the homogeneity hypothesis, Bartoo and Puri (1967) demonstrated (assuming the Poisson model) that it is more efficient to observe for relatively long periods on a few occasions than to allocate the same total observation time in shorter sessions over many occasions (e.g., four one-hour observations are more efficient than eight half-hour observations). A similar conclusion is indicated for Bernoulli-trial data and the test of $\sigma^2_\pi = 0$. Wisniewski (1972) demonstrated (assuming the binomial model) "that a few large samples are preferable to many small ones for detecting heterogeneity" (p. 680) (e.g., $T = 10$, $n_i = 10$ is better than $T = 20$, $n_i = 5$).

Consistency of individual differences. Currently available guidance on the design of observation schedules for addressing *Question B* is restricted to results that depend on the assumption of perfect consistency over time for each individual. The most extensive investigations of the efficiency of different observation schedules are found in Rowley (1976, 1978), where the effects of different observation schedules on reliability or generalizability coefficients are analyzed.[14] Rowley bases his analysis on the formulation of Ebel (1951).

---

[14] Less extensive studies with a similar orientation are Rosenshine's (1973) effort "to explore the question of the number of observations necessary to obtain a trustworthy sample of classroom transactions" (p. 221), Erlich and Shavelson's (1978) determination that "an unreasonable number of raters and occasions are required to measure certain variables reliably" (p. 88), and Erlich and Borich's (1979) analyses "concerning the number of observation occasions required to reach a 0.7 level of generalizability . . . for the case in which raters are well trained and not considered to be a significant source of error" (p. 11). Shavelson and Webb (1981, Section 2.6) advocate the use of multivariate generalizability analysis to determine the "optimal length of the observation period while taking into account the correlations among observation intervals" (p. 154).

Rowley's major finding is "for fixed total observation time, greater reliability is achieved by the use of a larger number of shorter, independent observations" (1978, p. 170). This general conclusion is documented in Figure 2 of Rowley (1978).

Interestingly, Rowley's conclusion about efficient schedules contradicts the statistical results cited above with regard to the consistency of an individual's behavior over time (i.e., testing and estimation for σ^2_λ and σ^2_π). This contradiction does not diminish the accuracy of either finding about the construction of efficient observation schedules. However, the contradiction does reinforce the commonsense notion that different designs will be optimal or desirable for different questions. Furthermore, recall that the conclusions of Rowley are based on the model of Ebel (1951) whose formulation employs the strong assumption that all individuals are consistent over time in their behavior (i.e., for each individual the homogeneity hypothesis is assumed to be satisfied). Thus Rowley's conclusions actually pertain to observation schedules for addressing *Question B* conditional on *Question A* being answered affirmatively for each individual. Therefore, Rowley's conclusions are not necessarily applicable for assessing consistency of individual differences using those statistical procedures based on the configuration of the individual time paths described in the previous section on indices of tracking.

Observation Instruments

As part of the design of a study of stability, the investigator must decide which target behaviors to observe and what information to collect on each target behavior. For behavior-count data the only information collected is the number of occurrences of the target behavior. However, this information is only a

fragment of the complete behavioral record; no record is obtained of the duration of the target behavior and the time elapsed between occurrences of the target behavior. Recording only the frequency of incidence of the target behavior, as is done in most classroom observation instruments, precludes many detailed statistical analyses of teacher behaviors and, perhaps most importantly for this research, precludes assessment of the validity of the assumptions (e.g., independence of events, distributional forms) underlying the statistical methods used to assess stability (see Appendix A). Similarly, for Bernoulli-trial data the sequence of outcomes of the individual trials contains valuable information that is lost when only the X_i and n_i are recorded.

Homogeneous Contexts

A key ingredient in studies of stability is designing the study so that a focused research question is addressed. Observing teacher behavior in homogeneous contexts is a basic requirement of a focused research question. Certainly, studies of stability that collect observations in as constant an environment as possible (e.g., group mathematics instruction) should precede studies that deliberately confound temporal and contextual facets (e.g., combining observations on both mathematics and English instruction over occasions). (It would seem unreasonable to expect teachers to be consistent over such disparate subject-matter contexts.) An unavoidable confounding occurs in studies of year-to-year temporal stability--the group of children in the teacher's class changes with the school year.

Conclusion

Previous empirical studies of stability of teacher behavior have been limited and unclear. The major weakness in these studies is the lack of an

explicit definition of stability of teacher behavior. Naturally, an attribute cannot be assessed without first being adequately defined.

Perhaps the most important contribution of this paper has been simply to formulate basic research questions about stability (i.e., *Questions A* and *B*). By linking these questions about stability to statistical models for various types of classroom observation data, we identify the statistical hypotheses and parameters that represent the consistency of teaching behavior over time or across contexts. To complete the development, statistical methods for the study of stability that follow from these representations are presented and illustrated.

A most striking consequence of the confusion and ambiguity in research on stability of teacher behavior is the absence of research on the consistency of the behavior of individual teachers. A concern with the consistency of the behavior of individual teachers is seen as far back as the writing of Barr (1929, especially p. 29). The methods we present should facilitate empirical research on the consistency of the behavior of individual teachers and perhaps serve the broader purpose of stimulating development of methods for addressing other research questions concerning the activities of individual teachers.

In closing, it is useful to consider what can be gained from this paper's contributions to the study of stability of teacher behavior. At the least, this paper ties together and demystifies the empirical and methodological literature on stability of teacher behavior. At the most, this paper may indicate important directions for research on teaching through a better understanding of and better methods for the study of the consistency of teaching behavior. In 1970, Flanders and Simon wrote in the Encyclopedia of Educational Research:

> the cutting edge of research on teaching effectiveness during the next decade may be more concerned with variation of teaching behavior between visits and with the consequences of this variation compared with the thrust of research that existed [in 1962] when the Gage Handbook went to press. (Flanders & Simon, 1970, p. 1425)

For better or worse, their prediccion has not been realized. We cannot judge how important the study of stability of teacher behavior will ultimately be for teaching effectiveness research. Yet, regardless of whether or not research on the "variation" of teaching behavior is to be prominent in research on teaching, it's a good idea to get it right.

References

Allington, R. L. (1980). Teacher interruption behaviors during primary-grade oral reading. Journal of Educational Psychology, 72, 371-377.

Altham, P. M. E. (1978). Two generalizations of the binomial distribution. Applied Statistics, 27, 162-167.

Armitage, P. (1955). Tests for linear trend in proportions and frequencies. Biometrics, 11, 375-386.

Barr, A. S. (1929). Characteristic differences in the teaching performance of good and poor teachers of the social studies. Bloomington, IL: Public School Publishing Co.

Bartoo, J. B., & Puri, P. S. (1967). On optimal asymptotic tests of composite statistical hypotheses. Annals of Mathematical Statistics, 38, 1845-1852.

Bayley, N. (1949). Consistency and variablity in the growth of intelligence for birth to eighteen years. Journal of Genetic Psychology, 75, 165-196.

Bennett, B. M., & Birch, J. B. (1974). On the small sample distribution and power of the log-likelihood ratio and variance tests for the Poisson. Journal of Statistical Computation and Simulation, 3, 33-40.

Bennett, B. M., & Kaneshiro, C. (1978). Small sample distribution and power of the binomial index of dispersion and log likelihood ratio tests. Biometric Journal, 20, 485-493.

Berliner, D. C. (1976). Impediments to the study of teacher effectiveness. Journal of Teacher Education, 27(1), 5-13.

Bliss, C. I. (1953). Fitting the negative binomial distribution to biological data. Biometrics, 9, 176-196.

Borich, G. D. (1977). Sources of invalidity in measuring classroom behavior. Instructional Science, 6, 283-318.

Bortkiewicz, L. V. (1931). The relations between stability and homogeneity. Annals of Mathematical Statistics, 2, 1-22.

Brophy, J. E. (1979). Teacher behavior and its effects. Journal of Educational Psychology, 71, 733-750.

Brophy, J. E., Coulter, C. L., Crawford, W. J., Evertson, C. M., & King, C. E. (1975). Classroom observations scales: Stability across time and context and relations with student learning gains. Journal of Educational Psychology, 67, 873-881.

Brophy, J. E., & Evertson, C. M. (1981). Student characteristics and teaching. New York: Longman.

Buhler, W., Fein, H., Goldsmith, D., Neyman, J., & Puri, P. S. (1965). Locally optimal test for homogeneity with respect to very rare events. Proceedings of the National Academy of Science, Mathematics, 54, 673-680.

Cox, D. R. (1955). Some statistical methods connected with series of events. Journal of the Royal Statistical Society (Series B), 17, 129-164.

Cox, D. R., & Isham, V. (1980). Point processes. New York: Chapman and Hall.

Cox, D. R., & Lewis, P. A. W. (1966). The statistical analysis of series of events. London: Chapman and Hall.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. Review of Educational Research, 53, 285-328.

Darwin, J. H. (1957). The power of the Poisson index of dispersion. Biometrika, 44, 286-289.

Detre, K., & White, C. (1970). The comparison of two Poisson-distributed observations. Biometrics, 26, 851-854.

Doyle, W. (1977). Paradigms for research on teacher effectiveness. In L. S. Shulman (Ed.), Review of research in education (Vol. 5). Itasca, IL: F. E. Peacock.

Ebel, R. L. (1951). Estimation of the reliability of ratings. Psychometrika, 16, 407-424.

Erlich, O., & Borich, G. (1979). Occurrence and generalizability of scores on a classroom interaction instrument. Journal of Educational Measurement, 16, 11-18.

Erlich, O., & Shavelson, R. J. (1978). The search for correlations between measures of teacher behavior and student achievement: Measurement problem, conceptualization problem, or both? Journal of Educational Measurement, 15, 77-89.

Evertson, C. M., Anderson, L. M., Edgar, D. P., Minter, M. D., & Brophy, J. E. (1977, May). Investigations of stability in junior high school math and English classes: The Texas Junior High School Study (R&D Report No. 4051; formerly No. 77-3). Austin, TX: Research and Development Center for Teacher Education, University of Texas at Austin.

Evertson, C. M., & Veldman, D. J. (1981). Changes over time in process measures of classroom behavior. Journal of Educational Psychology, 73, 156-163.

Fisher, R. A. (1950). The significance of deviations from expectation in a Poisson series. Biometrics, 6, 17-24.

Fisher, R. A. (1953). Note on the efficient fitting of the negative binomial. Biometrics, 9, 197-199.

Fisher, R. A., Thornton, H. G., & Mackenzie, W. A. (1922). The accuracy of the plating method of estimating the density of bacterial populations. Annals of Applied Biology, 9, 325-359.

Flanders, N. A. (1969). Teacher influence patterns and pupil achievement in the second, fourth, and sixth grade levels. Ann Arbor, MI: (ERIC Document Reproduction Service ED 051 123).

Flanders, N. A. (1970). Analyzing teaching behavior. Reading, MA: Addison-Wesley.

Flanders, N. A., & Simon, A. (1970) Teaching effectiveness: A review of the research, 1960-1966. In R. L. Ebel (Ed.), Encyclopedia of educational research. Chicago, IL: Rand McNally.

Forsyth, C. H. (1932). Proposal of a coefficient of stability. Journal of the American Statistical Association, 27, 173-176.

Forsyth, C. H. (1937). Some simple developments in the use of the coefficient of stability. Annals of Mathematical Statistics, 8, 5-11.

Foulkes, M. A., & Davis, C. E. (1981). An index of tracking for longitudinal data. Biometrics, 37, 439-446.

Gall, M. D. (1970). The use of questions in teaching. Review of Educational Research, 40, 707-721.

Gart, J. J. (1970). Some simple graphically oriented statistical methods for discrete data. In J. P. Patil (Ed.), Random counts in models and structures. University Park, PA: Pennsylvania State University Press.

Gbur, E. E. (1981). On the Poisson index of dispersion. Communications in Statistics: Simulation and Computation, 10, 531-535.

Good, T. L., Cooper, H. M., & Blakey, S. L. (1980). Classroom interactions as a function of teacher expectations, student sex, and time of year. Journal of Educational Psychology, 72, 378-385.

Haggard, E. A. (1958). Intraclass correlation and the analysis of variance. New York: Dryden Press, 1958.

Hendricks, W. A. (1935). A problem involving the Lexis theory of dispersion. Annals of Mathematical Statistics, 6, 78-82.

Heyde, C., & Seneta, E. E. (1977). Bienayme: Statistical theory anticipated. New York: Springer-Verlag.

Hoel, P. G. (1943). On indices of dispersion. Annals of Mathematical Statistics, 14, 155-162.

Johnson, N. L., & Kotz, S. (1970). Continuous univariate distributions--1. New York: Wiley.

Kendall, M. G., & Stuart, A. (1969). Advanced theory of statistics (Vol. 1). London: Griffin.

Kleinman, J. C. (1973). Proportions with extraneous variance: Single and independent samples. Journal of the American Statistical Association, 63, 46-54.

Klotz, J. (1973). Statistical inference in Bernoulli trials with dependence. Annals of Statistics, 1, 373-379.

Lomax, R. G. (1982). An application of generalizability theory to observational research. Journal of Experimental Education, 51, 22-30.

McGaw, B., Wardrop, J. L., & Bunda, M. A. (1972). Classroom observation schemes: Where are the errors? American Educational Research Journal, 9, 13-27.

McMahan, C. A. (1981). An index of tracking. Biometrics, 37, 447-455.

Medley, D. M. (1979). The effectiveness of teachers. In P. L. Peterson and H. J. Walberg (Eds.), Research on teaching: Concepts, findings, and interpretations. Berkeley, CA: McCutchan.

Medley, D. M. (1982). Teacher effectiveness. In H. E. Mitzel (Ed.), Encyclopedia of Educational Research (5th ed.). Riverside, NJ: Macmillan.

Medley, D. M., & Mitzel, H. E. (1963). Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago, IL: Rand McNally.

Moon, T. C. (1969). A study of verbal behavior patterns in primary grade classrooms during science activities. Doctoral dissertation, College of Education, Michigan State University.

Moon, T. C. (1971). A study of verbal behavior patterns in primary grade classrooms during science activities. Journal of Research in Science Teaching, 8, 171-177.

Paul, S. R., & Plackett, R. L. (1978). Inference sensitivity for Poisson mixtures. Biometrika, 65(3), 591-602.

Potthoff, R. F., & Whittinghill, M. (1966). Testing for homogeneity. I. The binomial and multinomial distributions. Biometrika, 53, 167-182. (a)

Potthoff, R. F., & Whittinghill, M. (1966). Testing for homogeneity. II. The Poisson distribution. Biometrika, 53, 183-190. (b)

Redfield, D. L., & Rousseau, E. W. (1981). A meta-analysis of experimental research on teacher questioning behavior. Review of Educational Research, 51, 237-245.

Robertson, A. (1951). The analysis of heterogeneity in the binomial distribution. Annals of Eugenics, 16, 1-15.

Rogosa, D. R., & Willett, J. B. (1983). Comparing two indices of tracking. Biometrics, in press.

Rosenshine, B. (1971). Teaching behaviors and student achievement. London: National Foundation for Educational Research.

Rosenshine, B. (1973). The smallest meaningful sample of classroom transactions. Journal of Research in Science Teaching, 10, 221-226.

Rosenshine, B., & Furst, N. (1973). The use of direct observation to study teaching. In R. M. W. Travers (Ed.), Second handbook of research on teaching. Chicago, IL: Rand McNally.

Rowley, G. L. (1976). The reliability of observational measures. American Educational Research Journal, 13, 51-59.

Rowley, G. L. (1978). The relationship of reliability in classroom research to the amount of observation: An extension of the Spearman-Brown formula. Journal of Educational Measurement, 15, 165-180.

Ryan, F. L. (1973). Differentiated effects of levels of questioning on student achievement. Journal of Experimental Education, 41, 63-67.

Ryan, F. L. (1974). The effects on social studies achievement of multiple student responding to different levels of questioning. Journal of Experimental Education, 42, 71-75.

Sandoval, J. (no date). The evaluation of teacher behavior through observation of videotape recordings. Beginning teaching evaluation study, Phase II (Volume III, 3 Final Report). Washington, DC: National Institute of Education, U.S. Department of Education.

Shavelson, R. J., Berliner, D. C., Ravitch, M. M., & Loeding, D. (1974). Effects of position and type of question on learning from prose material: Interaction of treatments with individual differences. Journal of Educational Psychology, 66, 40-48.

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. Review of Educational Research, 46, 553-611.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973-1980. British Journal of Mathematical and Statistical Psychology, 34, 133-166.

Shenton, L. R., & Myers, R. (1963). Comments on estimation for the Negative Binomial distribution. In G. P. Patil (Ed.), Classical and contagious discrete distributions. London: Pergamon.

Snedecor, G. W., & Cochran, W. G. (1980). Statistical methods (7th ed.). Ames, IA: Iowa State University Press.

Tarone, R. E. (1979). Testing the goodness of fit of the binomial distribution. Biometrika, 66, 585-590.

Trinchero, R. L. (1974). Three technical skills of teaching: Their stability and effect on pupil attitudes and achievement. Doctoral dissertation, Stanford University.

Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

Winne, P. H. (1979). Experiments relating teachers' use of higher cognitive questions to student achievement. Review of Educational Research, 49, 13-49.

Wisniewski, T. K. M. (1972). Power of tests of homogeneity of a bionomial series. Journal of the American Statistical Association, 67, 680-683.

Wohlwill, J. F. (1973). The study of behavior development. New York: Academic Press.

Wohlwill, J. F. (1980). Cognitive development in childhood. In O. G. Brim and J. Kagan (Eds.), Constancy and change in human development. Cambridge, MA: Harvard University Press.

APPENDIX A

Appendix A

Statistical Procedures

In this Appendix we present details of the statistical procedures for addressing *Question A* with Behavior-count or Bernoulli-trial data. For each type of data, we present statistical procedures for testing the homogeneity hypothesis and for estimating the amount of heterogeneity. In addition, the statistical models and the assumptions underlying these procedures are discussed. Numerous references to statistical literature on discrete distributions and point processes are provided to guide the reader to further treatments of relevant technical issues.

Homogeneity Hypothesis: Behavior-Count Data

Poisson Model

The statistical procedures for Behavior-count data are based on the natural assumption of a Poisson distribution for the counts. That is, on occasion i, the probability of X events occurring in an interval b_i is

$$\frac{e^{-b_i\lambda_i}(b_i\lambda_i)^X}{X!} .$$

For any single occasion the natural estimate of λ_i is the empirical rate of observed behavior at the occasion, $\hat{\lambda}_i = X_i/b_i$, and the estimate of μ_λ is the weighted average $\bar{\lambda} = \sum_i X_i / \sum_i b_i$ (the maximum likelihood estimate under the homogeneity hypothesis).

For each teacher, each occasion of observation is assumed to provide an independent sample of behavior; that is, the X_i are assumed to be independent across occasions. Under the homogeneity hypothesis ($\lambda_i = \lambda$ for all i) the

distribution of each X_i is Poisson with mean $b_i\lambda$, and the resulting model is the distribution function given above with the common λ replacing λ_i (see Potthoff and Whittinghill, 1966b, Equation 1).

Within an occasion the assumption of a Poisson distribution for the X_i is satisfied if the individual events are generated by a Poisson process. The Poisson process "plays a role in point process theory in most respects analagous to that of the normal distribution in the study of random variables" (Cox & Isham, 1980, p. 45).

Assessments of the validity of assumptions about the distribution of the X_i, or about the point process assumed to generate the X_i, require data on the individual events (such as waiting times between events). Cox and Lewis (1966, chapter 6) present a number of methods for testing general renewal process models and specifically, in section 6.3, tests for Poisson processes are presented. In particular, the validity of the assumption of a Poisson distribution within each occasion cannot be evaluated from just the X_i and b_i. In serious empirical work, assessments of the validity of the statistical model should be made. Marked deviations from the assumption of a Poisson distribution, such as those that may be introduced by severe dependence among individual events (see Cox and Lewis, 1966, chapter 2 and 7 for definitions of independence and non-independence in series of events), may render tests of the homogeneity hypothesis equivocal because positive dependence within occasions may not be distinguishable from heterogeneity across occasions.

Test Statistics

To test the null hypothesis $\lambda_i = \lambda$ for all i against the general alternative that not all the λ_i are equal we use the statistic

$$\sum_{i=1}^{T} \frac{\{X_i - (b_i/\Sigma_i b_i)\Sigma_i X_i\}^2}{(b_i/\Sigma_i b_i)\Sigma_i X_i} , \qquad \text{(A1)}$$

which is distributed approximately as χ^2 with T - 1 degrees of freedom under the null hypothesis (see Potthoff and Whittinghill, 1966b). Thus the homogeneity hypothesis is rejected, at level α , when the test statistic exceeds the critical value $\chi^2_{T-1}(\alpha)$. The test statistic in Expression A1 assesses whether the $\hat{\lambda}_i$ are more spread out (over occasions) than would be expected under the homogeneity hypothesis (given the Poisson model).

The structure of the test statistic may be understood more clearly from the alternative expression:

$$\frac{\sum_{i=1}^{T} b_i(\hat{\lambda}_i - \bar{\lambda})^2}{\bar{\lambda}} . \qquad \text{(A2)}$$

Expression (A2) shows that the numerator of the test statistic is a weighted variance of the $\hat{\lambda}_i$. The b_i , which are known to the analyst, are often fixed in advance by the observation schedule. Alternatively, the b_i may be determined by the immediate classroom situation (e.g., in the observation of teaching behavior during reading instruction in an elementary-school classroom, the length of observation depends on how long the teacher carries out reading instruction). See also Potthoff and Whittinghill (1966b, p. 183).

When all the b_i equal one, the test statistic in Expressions A1 and A2 reduces to the familiar "Poisson Index of Dispersion" (also known as the "variance test") introduced by R. A. Fisher (see Fisher, Thornton & Mackenzie, 1922; Fisher, 1950; Hoel, 1943). This statistic has the simple form

$$\frac{\sum_i (X_i - \bar{X})^2}{\bar{X}} \quad . \qquad \text{(A3)}$$

In Expression A3 the numerator is the sample variance (over occasions) multiplied by $T - 1$, and the denominator $\bar{X}$ is an estimate of the variance (within occasions) under the homogeneity hypothesis.

Potthoff and Whittinghill (1966b) showed that Expression A3 is the locally most powerful unbiased test against the negative binomial alternative (i.e., the λ_i follow a gamma distribution). Extensive study has shown the tests based on Expressions A3 and A1 to have reasonably good power against a variety of alternatives (see Bennett and Birch, 1974: Darwin, 1957; Gbur, 1981; Paul and Plackett, 1978).

Alternative Statistics. A likelihood ratio statistic for testing this homogeneity hypothesis has been presented by Cox and Lewis (1966, Section 9.3, Equation 8). This statistic is asymptotically equivalent to Expression A1 and yields nearly identical results to Expression A1 in small samples. Other statistics for testing the homogeneity hypothesis are designed to be sensitive to the alternative that the λ_i follow a gamma distribution as opposed to the null hypothesis that $\lambda_i = \lambda$. (The use of the gamma distribution for the λ_i is for mathematical convenience because it yields a negative binomial distribution for the X_i.) Test statistics designed to be optimal for the gamma alternative are examined in Potthoff and Whittinghill (1966b) and Buhler, et al. (1965). For applications, the test statistic in Expression A1 should be used, unless strong reasons exist for positing a gamma distribution for the λ_i.

Estimating Heterogeneity: Behavior-Count Data

The variance of the distribution of the λ_i, σ_λ^2, represents the heterogeneity over occasions of the rate of the target behavior. In this paper

we use estimates of σ_λ^2 to describe the behavior of individual teachers. The statistical model for the estimation of heterogeneity is the same as that used for testing the homogeneity hypothesis; however, the estimates of heterogeneity are much less vulnerable to moderate violations of the statistical model and its assumptions.

Variance Component Estimates. Cox (1955, Section 5.3) developed useful estimates of σ_λ^2, the simplest of which is:,

$$\frac{(d - 1)(T - 1)\bar{\lambda}}{\sum_i b_i - \sum_i b_i^2 / \sum_i b_i} \quad , \tag{A4}$$

where d is the test statistic in Expression A1 divided by $T - 1$. The estimate for σ_λ^2 used in the analyses presented in this paper is an adaptive estimator related to Expression A4 (for details see Cox, 1955, Section 5.3). Very small values of the test statistic in Expression A1 may result in corresponding estimates of σ_λ^2 that are negative. In such cases the estimate of σ_λ^2 is set to zero, as is consistent with a failure to reject the homogeneity hypothesis.

Negative Binomial Estimates. When all the $b_i = 1$ and an assumption of a gamma distribution for the λ_i can be made, estimation of the variance of the gamma distribution is based on estimation for the resulting negative binomial distribution of the X_i. Hence, a method of moments estimate for σ_λ^2 (termed the "Evans-Anscombe" estimate by Shenton and Myers, 1963) is simply

$$\frac{\sum_i (X_i - \bar{X})^2}{T} - \bar{X} \quad . \tag{A5}$$

The maximum likelihood estimate for σ_λ^2 under these assumptions was developed by Fisher (1950, 1953; see also Bliss, 1953). Although iterative methods are

required, the computation of the maximum likelihood estimate is straightforward (see Johnson and Kotz, 1969, Section 5.8).

Homogeneity Hypothesis: Bernoulli-Trial Data

Binomial Model

The statistical procedures for Bernoulli-trial data are based on the assumption that the sum of the outcomes of the Bernoulli-trials on any single occasion follow a binomial distribution. That is, for occasion i the probability that X successes occur in the n_i trials is

$$\binom{n_i}{X} (\pi_i)^X (1 - \pi_i)^{n_i - X} \quad .$$

For any single occasion the natural estimate of π_i is the empirical proportion of successes, $p_i = X_i/n_i$, and the estimate of μ_π is the weighted average $\bar{p} = \sum_i X_i / \sum_i n_i$ (the maximum likelihood estimate under the homogeneity hypothesis).

For each teacher, each occasion of observation is assumed to provide an independent sample of behavior, yielding, for each teacher, T independent samples of sizes n_1 , n_2 ,..., n_T . Under the homogeneity hypothesis ($\pi_i = \pi$ for all i) the distribution of each X_i is binomial with the same parameter π, and the resulting model is the distribution function given above with the common π replacing π_i (see Potthoff and Whittinghill, 1966a, Equation 1).

Within an occasion, the assumption of a binomial distribution for the sum of the Bernoulli trails is satisfied if the Bernoulli-trials are identically and independently distributed (i.i.d.). Assessments of the validity of the assumptions about the structure of individual events require data on the individual events. That is, the validity of the assumption of a binomial distribution within each occasion cannot be evaluated from just the X_i and n_i.

Various models for dependence among the Bernoulli-trials are studied by Altham (1978) and Klotz (1973); Klotz (1973, Section 6) and Tarone (1979, Section 2) developed estimation and testing procedures for detecting dependence of the Bernoulli-trials within occasions. Violations of the assumption of i.i.d. Bernoulli-trials are important for the statistical procedures we use only insofar as the assumption of a binomial distribution for the X_i is undermined (especially with regard to the variance of X_i being $n_i\pi_i(1 - \pi_i)$). Minor violations of the assumption of i.i.d. Bernoulli-trials will not greatly affect even the statistical tests of the homogeneity hypothesis. However, in serious empirical work, assessments of the validity of assumption of i.i.d. Bernoulli trials should be made. Marked deviations from the assumption of a binomial distribution, such as those that may be introduced by severe dependence amongst the Bernoulli-trials, may render tests of the homogeneity hypothesis equivocal, as positive dependence within occasions may not be distinguishable from heterogeneity across occasions.

Test Statistics

To test the null hypothesis that $\pi_i = \pi$ for all i against the general alternative that not all the π_i are equal we use the "Binomial index of dispersion"

$$\sum_{i=1}^{T} \frac{(X_i - n_i\bar{p})^2}{n_i\bar{p}(1 - \bar{p})} , \qquad \text{(A6)}$$

which is distributed approximately (for n_i not small) as X^2 with $T - 1$ degrees of freedom under the null hypothesis (see Hoel, 1943; Potthoff & Whittinghill, 1966a; Wisniewski, 1968, 1972). (A familiar use for this statistic is in

testing for "independence" in a $2 \times T$ contingency table; see, for example, Snedecor and Cochran, 1980, Section 11.7). Thus the homogeneity hypothesis is rejected, at level α, when this test statistic exceeds the critical value $\chi^2_{T-1}(\alpha)$. This test statistic assesses whether the p_i are more spread out (over occasions) than would be expected under the homogeneity hypothesis (given the binomial model). An interesting historical note is that the test statistic in Expression A6 divided by its degrees of freedom is the Lexis quotient, which was prominent in the late nineteenth and early twentieth centuries in the study of consistency or stability of statistical series (see Bortkiewicz, 1931; Forsyth, 1932, 1937; Heyde and Seneta, 1977; Lexis, 1877).

For the special case of $n_i = n$ Potthoff and Whittinghill (1966a) and Gart (1970) demonstrated that the statistic in Expression A6 is optimal, in the sense of locally most powerful unbiased, against the beta-binomial alternative. Power functions for the test based on Expression A6 for a variety of alternative distributions have been studied by Bennett and Kaneshiro (1978) and Wisniewski (1972)

Alternative Statistics. Another statistic for testing the homogeneity hypothesis is obtained from the likelihood ratio criterion (Bennett & Kaneshiro, 1978, Equation 4). This test statistic is asymptotically equivalent to Expression A6, and the numerical results of Bennett and Kaneshiro (1978) show that small sample properties favor the use of Expression A6. Other statistics for testing the homogeneity hypothesis are designed to be sensitive to the alternative that the π_i follow a beta distribution as opposed to the null hypothesis that $\pi_i = \pi$. The use of the beta distribution is for mathematical convenience as it yields a beta-binomial distribution for the X_i. An

asymptotically optimal test statistic for the beta-binomial alternative is derived in Section 3 of Tarone (1979); see also Gart (1970), Potthoff and Whittinghill (1966a) and Wisniewski (1968). For applications, Expression A6 should be used unless strong reasons exist for positing the beta distribution for the π_i .

Estimating Heterogeneity: Bernoulli-Trial Data

The variance of the distribution of the π_i , σ_π^2 , represents the heterogeneity over occasions for a behavior having the form of a Bernoulli trial. In this paper we use estimates of σ_π^2 to describe the behavior of individual teachers. The estimation of heterogeneity relies on the binomial model and its assumptions; however, the estimates of heterogeneity are much less vulnerable than the test of the homogeneity hypothesis to moderate violations of the model.

Three estimates of σ_π^2 have been developed in the statistical literature. The simplest estimate, developed by Hendricks (1935), is

$$\{\sum_i (p_i - \bar{p})^2 - \bar{p}(1 - \bar{p})\sum_i (1/n_i)\}/T \qquad \text{(A7)}$$

and can be derived by inverting the Lexis Formula. Another estimate of σ_π^2, obtained by Robertson (1951), is ,

$$\frac{\{d - (T - 1)\}\bar{p}(1 - \bar{p})}{\sum_i n_i - (\sum_i n_i^2 / \sum_i n_i) - (T - 1)} \quad , \qquad \text{(A8)}$$

where d is the statistic in Expression A6 divided by T - 1 . This estimate can be derived by solving for σ_π^2 in the expectation of the test statistic in Expression A6. When the value of the test statistic in Expression A6 is very small, either or both of Expressions A7 and A8 may produce negative estimates of

σ_π^2. In such cases the estimate of σ_π^2 should be set to zero, as is consistent with a failure to reject the homogeneity hypothesis.

Kleinman (1973) developed an iterative estimate of σ_π^2, based on method of moments estimation for the beta-binomial distribution (defined by Equations 2.5 through 2.8 in Kleinman, 1973). In practice, the simple estimate of Hendricks (Expression A7) agrees closely with Kleinman's estimate, whereas Robertson's estimate (Expression A8) is consistently larger than the other two. In the examples in the text, Kleinman's estimate is used. Kleinman's estimate is constrained to produce only non-negative estimates of σ_π^2.

Appendix B

Data Sources

Texas Junior High School Study

During the 1974-1975 school year, a large-scale process-product study at the junior-high level was conducted by the Research and Development Center for Teacher Education (Evertson & Veldman, 1981; Evertson et al., 1977). Data collection began in November 1974 and continued until mid-May 1975. Each of 68 seventh- and eight-grade teachers (29 math and 39 English) were observed in two of their classes. In each of the 136 classes about 20 one-hour observations were conducted, yielding data on a variety of high-inference (global rating) and low-inference (frequencies of behavior) measures.

In our use of these data the original 20 distinct occasions of observation have been combined by pooling observations within a calendar month into 6 different occasions (November through April). There is nothing superior or necessarily desirable about this grouping of the one-hour observations; our analyses use the grouped data only because this grouping was also present in the analyses of the Texas Junior High School data by Evertson and Veldman (1981), from whom we obtained these data. (Also, the data we obtained were in the form of monthly rates of behavior; we determined the X_i and b_i for the monthly data from the fractional rates.) Data from the Texas Junior High School Study are used in Tables 1 and 5 (the low-inference variable, Behavioral Criticism), Figure 3 and Table 6 (the low-inference variable, Product Questions), Table 4 (the high-inference variable, Positive Affect), and in the section Stability Across Contexts (the low-inference variables, Product Questions and Call-outs). In Tables 1, 4, and 5 we retain the original teacher-class identification codes from the Texas Junior High School Study. For example, the identification number 21052 in Table 1 denotes observations of English teacher number 05 in school 1 for the class scheduled during the second period of the school day.

Moon's Elementary-School Science Study

Moon (1969, 1971) studied the verbal behavior patterns of 32 mid-Michigan elementary-school teachers. Sixteen of these teachers used a new science curriculum (SCIS), while the other 16 "conventional" teachers served as a control group. Each of the 16 teachers of the SCIS curriculum was observed twice in the spring of 1968, before the summer workshop on the new curriculum, and on four occasions during the following school year. The control teachers were observed only twice.

Audio recordings of the science lessons at each of the six observation occasions were coded using both Flanders' interaction-analysis system and an independently developed instrument for counting and classifying teacher questions. Moon's data on teacher questions have been used to compute "stability" coefficients by Rosenshine (1973), who used the last four observations on the SCIS teachers, and by Shavelson and Dempsey-Atwood (1976), who used the observational data on both the SCIS and conventional teachers. Shavelson and Dempsey-Atwood also reported "stability coefficients" for the indirectness-directness (I/D) ratio obtained from Flanders' instrument.

From Moon's original coding sheets we were able to recover counts of each of the five teacher-question types for each of the six observation occasions. In Tables 2, 7, 8, 9, and 10 and in Footnote 12, the data on teachers' use of recall-facts questions (lower-order questions) on the last four observation occasions (those following the summer workshop) were used. Also, Table 3 presents five data points for Flanders' indirectness-directness (I/D) ratio. The first datum is the average of the indirectness-directness (I/D) ratios for the two pre-workshop observations; the remaining data points represent observations 3 through 6.

Trinchero's Student Teacher Study

Trinchero (1974) used videotapes of English and social-studies student teachers for his dissertation study. The tapes were made as part of the Stanford Teacher Education Program during the 1967-1968 school year. Each student teacher taught pre-set lessons to a group of 9th and 10th graders, who were paid volunteers. In 1972 Trinchero had observers code these tapes, recording counts for different categories of teacher questions for three occasions of observation. Using data for two occasions, Shavelson and Dempsey-Atwood (1976, Table 1) report "stability coefficients" for a variety of these teacher questioning behaviors. The data on teacher questions provided the Bernoulli-trial data for lower-order questions used in Table 10.